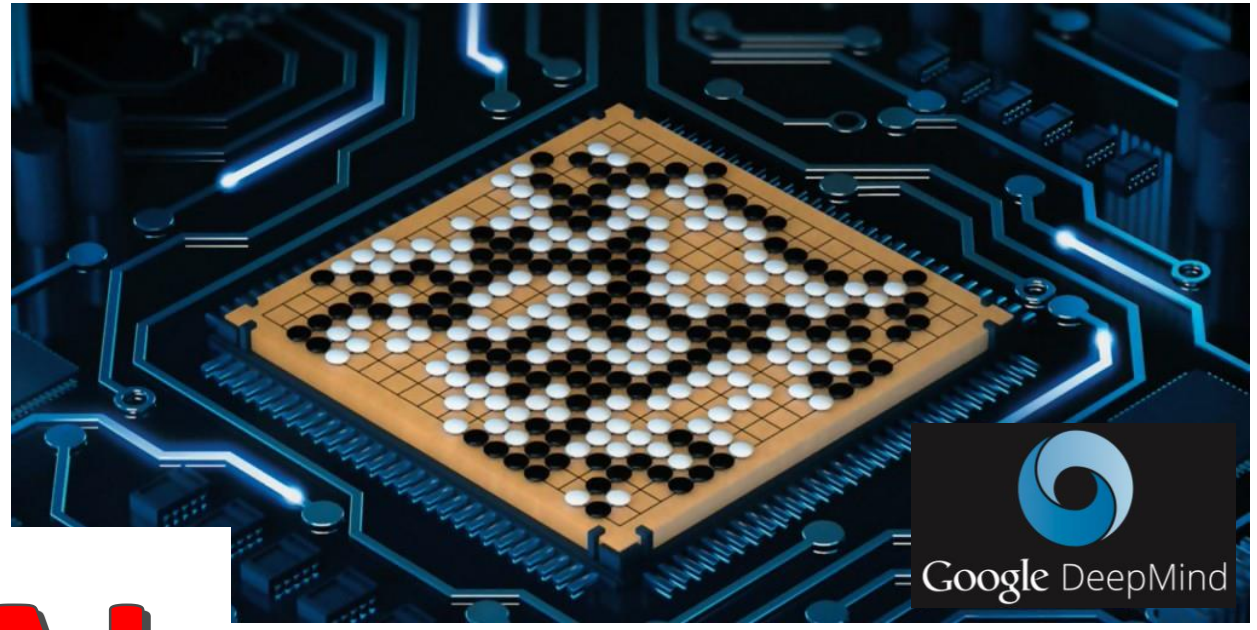
Trustworthy Human-Centric AI

Fredrik Heintz, Dept. of Computer Science
Linköpings universitet
fredrik.heintz@liu.se
@FredrikHeintz

LINKÖPINGS UNIVERSITET

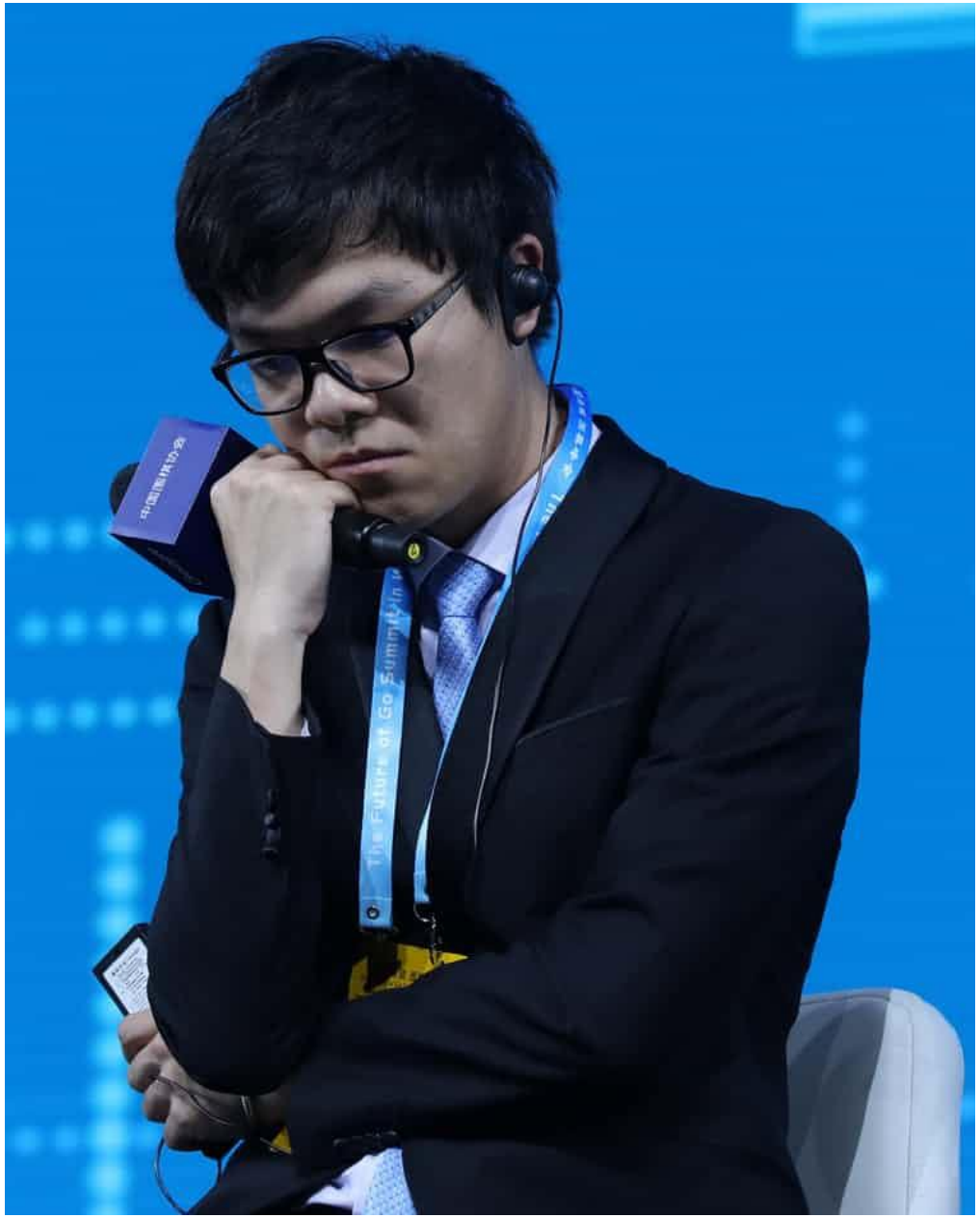
SWEDISH ARTIFICIAL INTELLIGENCE SOCIETY


Google DeepMind

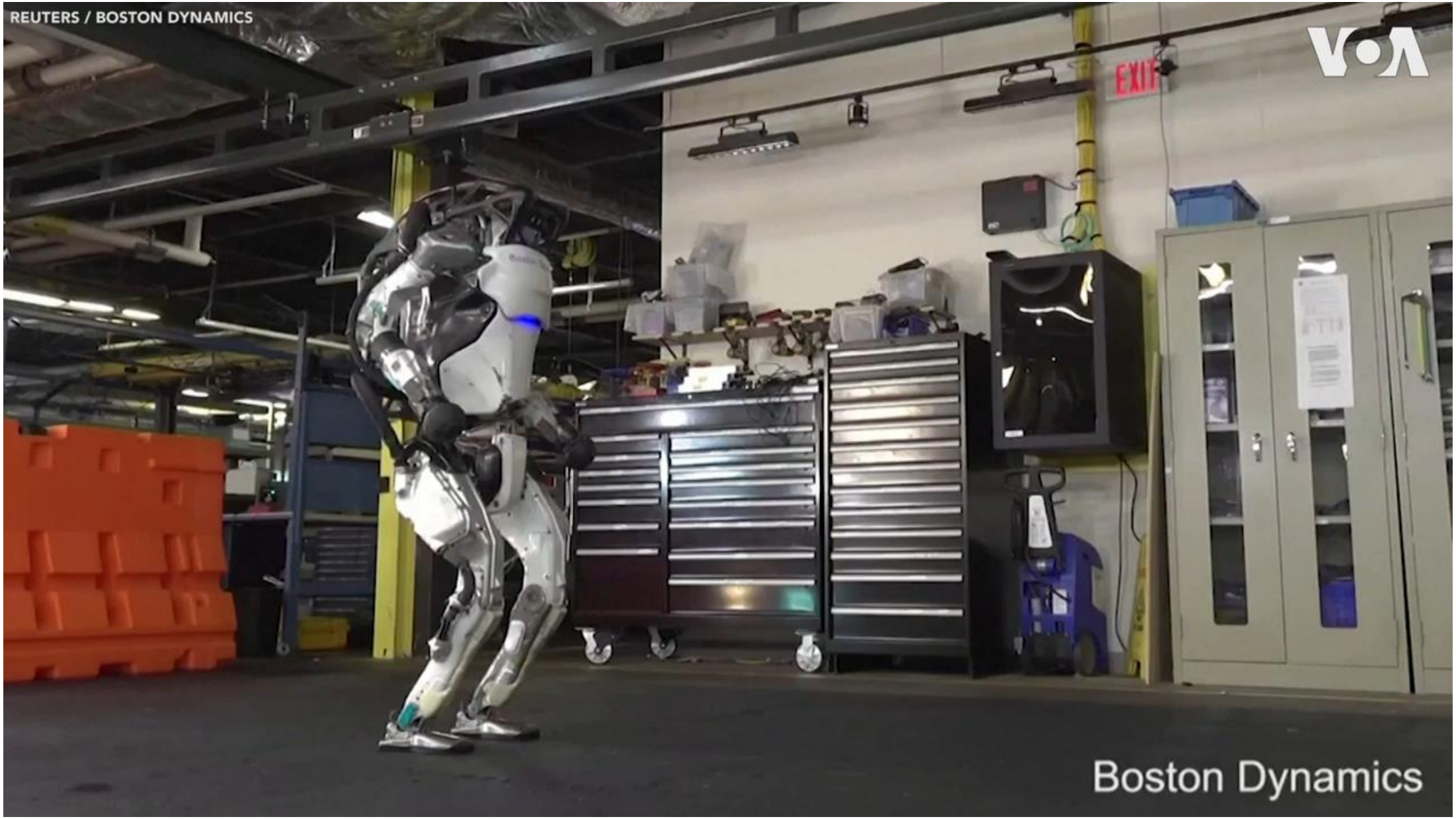
AI


THINK
$300,000
$1,000,000
$200,000
KEN
WATSON
BRAD


BostonDynamics
Boston Dynamics




柯洁 KE JIE
00:46:57



LINKÖPINGS UNIVERSITET

https://www.youtube.com/watch?v=YdnJI9T-yXI

SWEDISH ARTIFICIAL INTELLIGENCE SOCIETY

https://www.youtube.com/watch?v=W0_DPi0PmF0

LINKÖPINGS UNIVERSITET

https://www.thispersondoesnotexist.com/

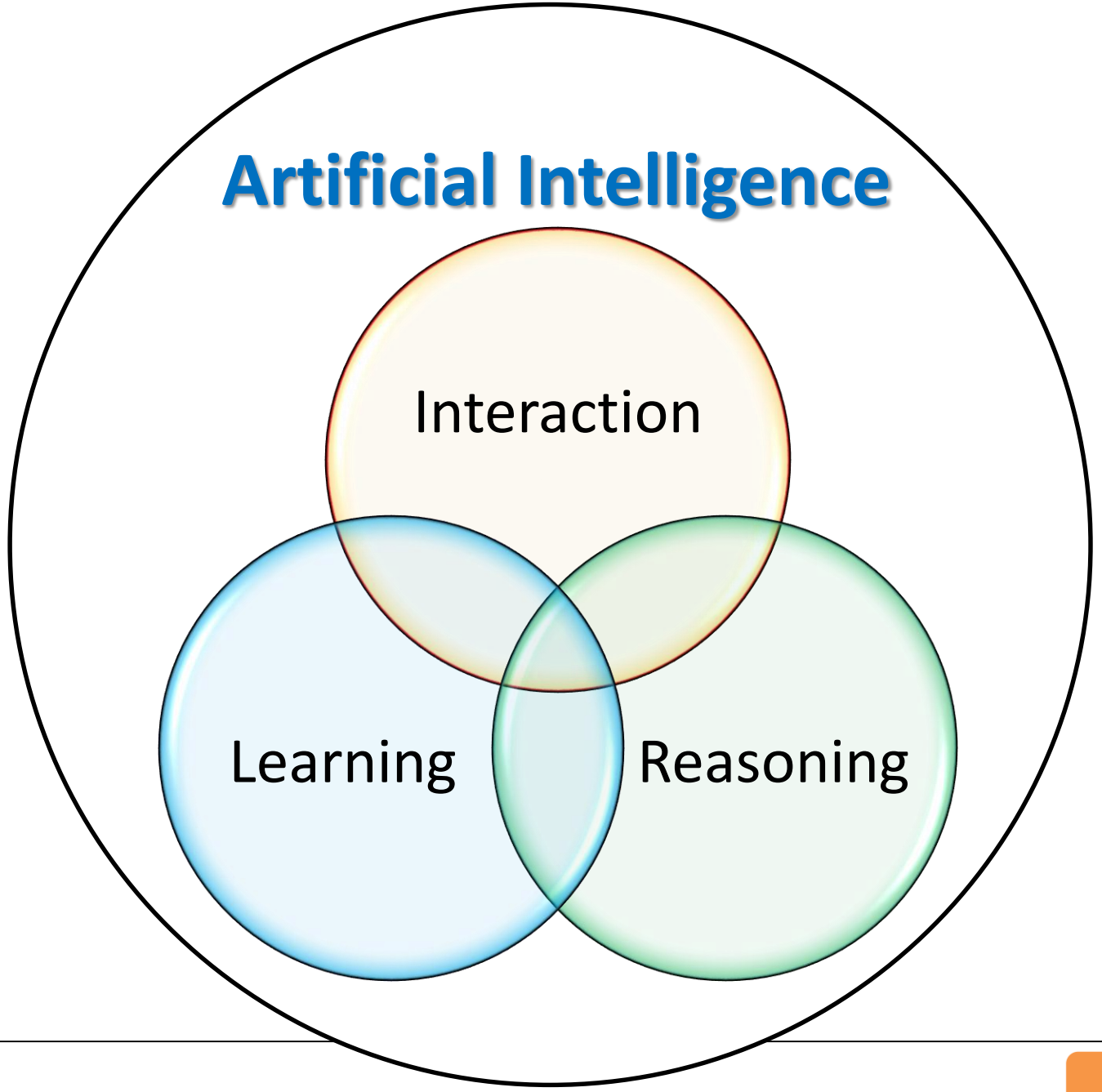
SWEDISH ARTIFICIAL INTELLIGENCE SOCIETY

Artificial Intelligence – What is it? – Definitions

"Artificial Intelligence is the **science** and **engineering** of **making intelligent machines**, especially intelligent computer programs."
- John McCarthy, Stanford

"Artificial intelligence (AI) refers to **systems that display intelligent behaviour** by analysing their environment and taking actions – with some degree of **autonomy** – to achieve specific **goals**."
- EU Communication 25 April 2018

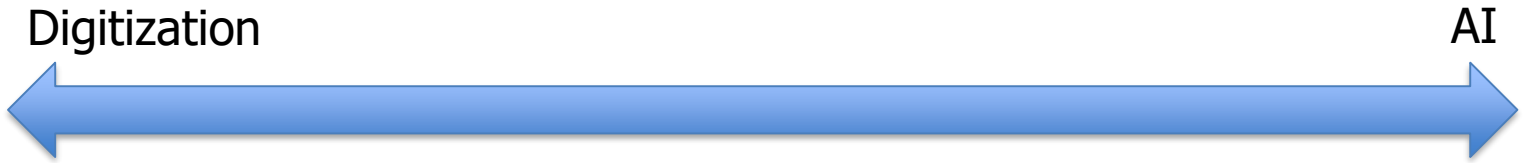
"the scientific understanding of the **mechanisms underlying thought** and **intelligent behavior** and their embodiment **in machines**." - AAAI


Artificial Intelligence
Interaction
Learning
Reasoning

Digitalization and AI

Digitization → **Big Data** → **AI**

first wave | second wave | third wave



Well defined problems
Predictable situations
Structured data
General solutions
Rationalizes
Evolutionary
…

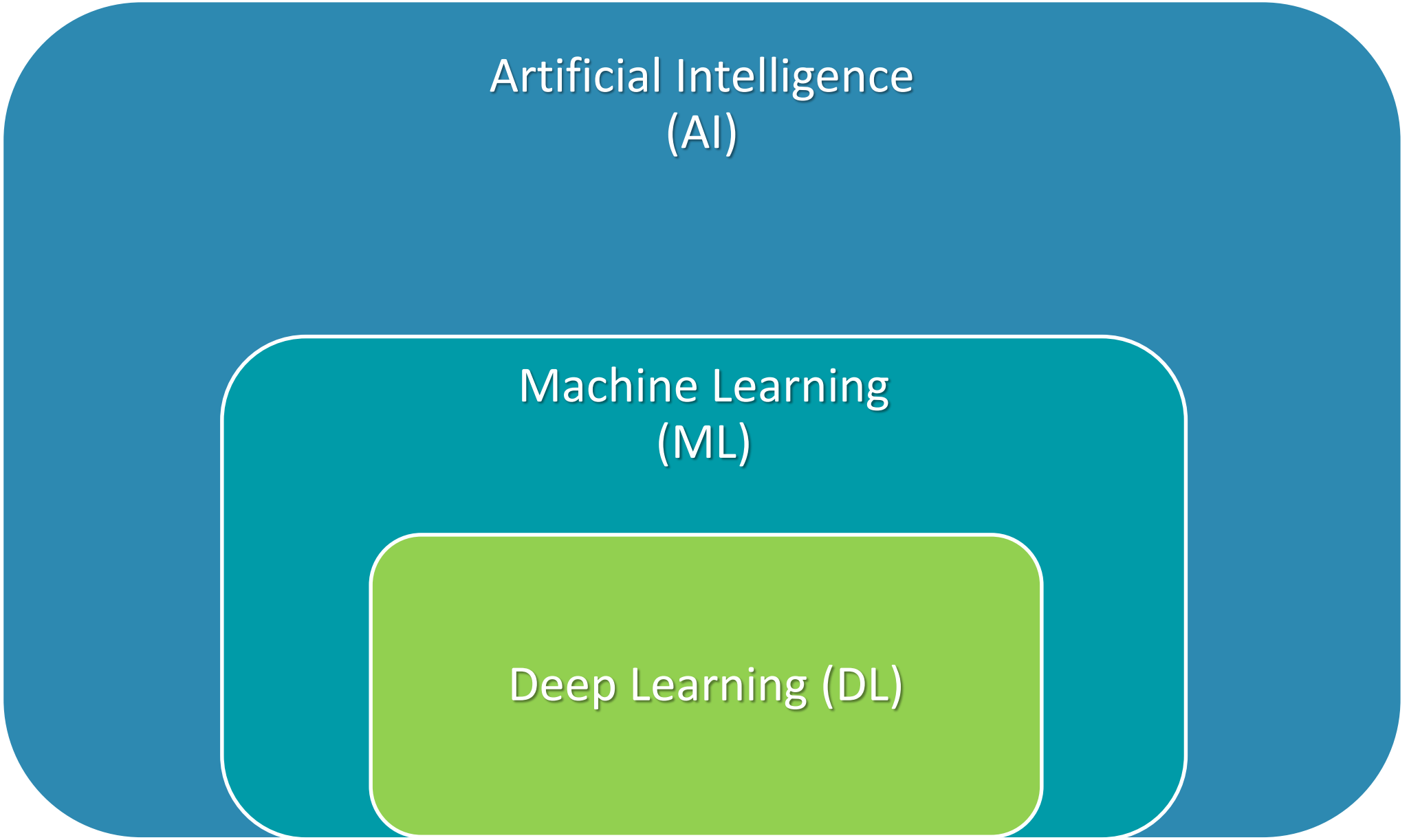
Hard to define problems
Unanticipated situations
Unstructured data
Adaptable solutions
Amplifies
Revolutionary
…

Skolans Digitalisering

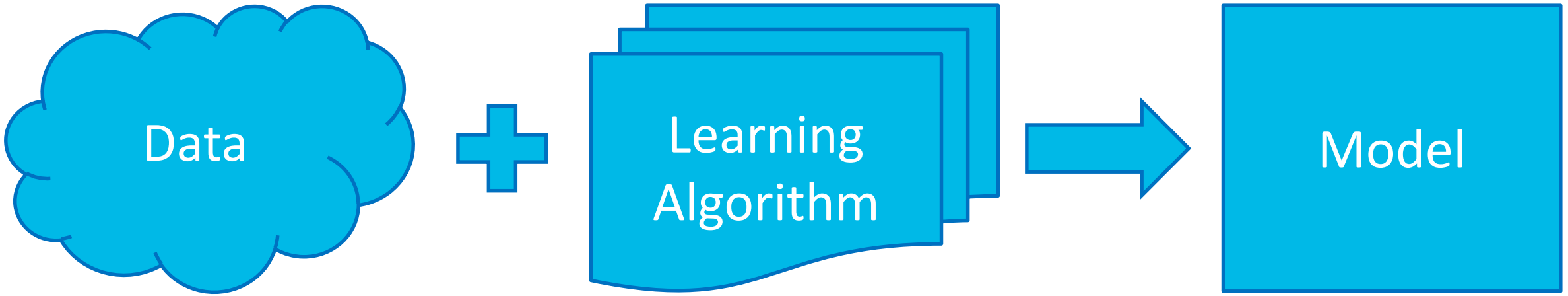
A. *Skolans digitala infrastruktur*, tillgång till datorer, fungerande nätverk, hantering av inköp av program, supportpersonal mm.
B. Innehållet i undervisningen för att möte de behov som dagens digitaliserade samhälle kräver, ***vad*** *ska undervisas*.
C. Digitalisering av skolans undervisning och pedagogik utifrån de digitala verktyg och möjligheter som finns idag, ***hur*** *undervisningen genomförs*.
D. *Digitaliseringen av skolans stödsystem* och övrig verksamhet, t.ex. administrativa system, interaktion med föräldrar, intranät, osv.

Hur kan man dra nytta av AI?

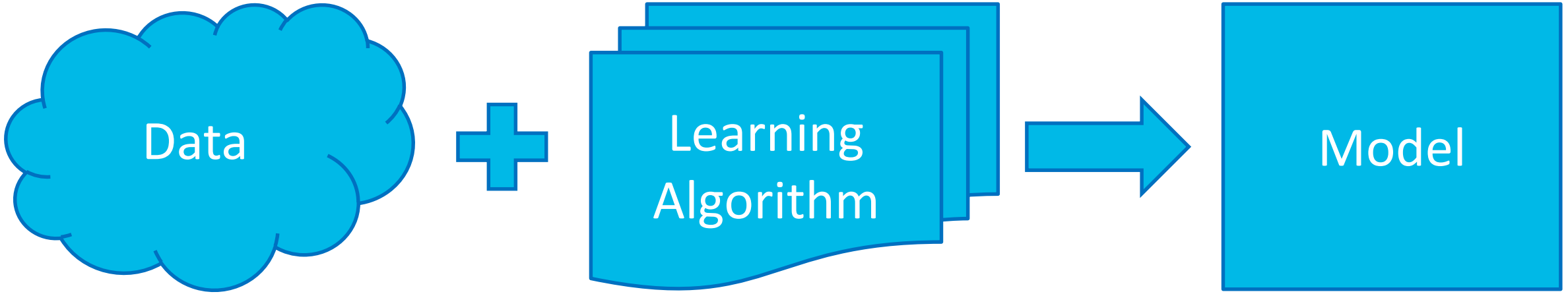
- Textbaserad AI – hantera stora mängder text och dokument
  - Sammanfatta, organisera, analysera text
  - Arbetsförmedlingens nya jobbmatchining
  - Nyhetsvärdering från diariet
- Dialogbaserad AI – hantera kommunikation
  - En ingång till kommunen/verksamheten
  - Bara fråga om information en gång
- Analytisk AI – förstå och förutspå vad som händer
  - Förstå faktisk användning av resurser
  - Förutspå kommande behov

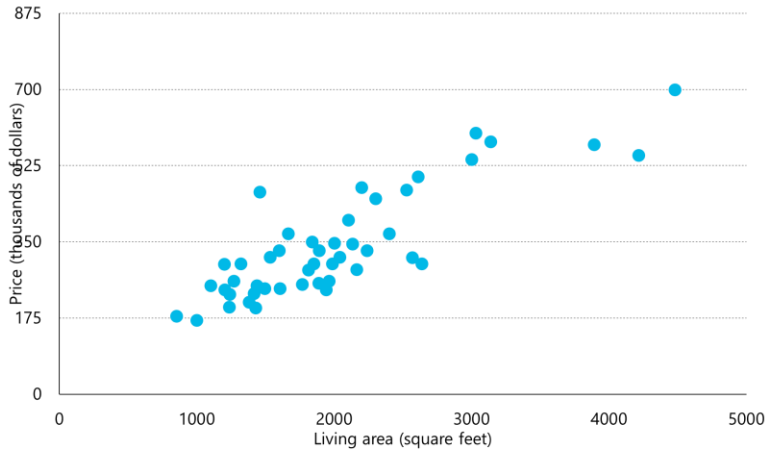

Artificial Intelligence
(AI)
Machine Learning
(ML)
Deep Learning (DL)

Machine Learning



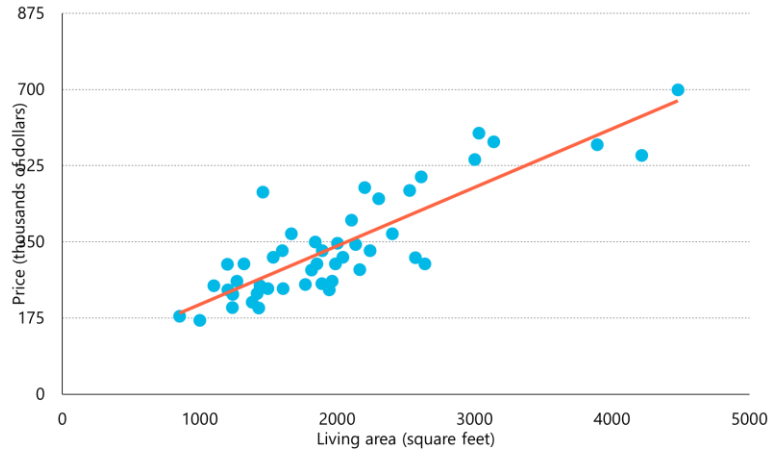
Machine Learning



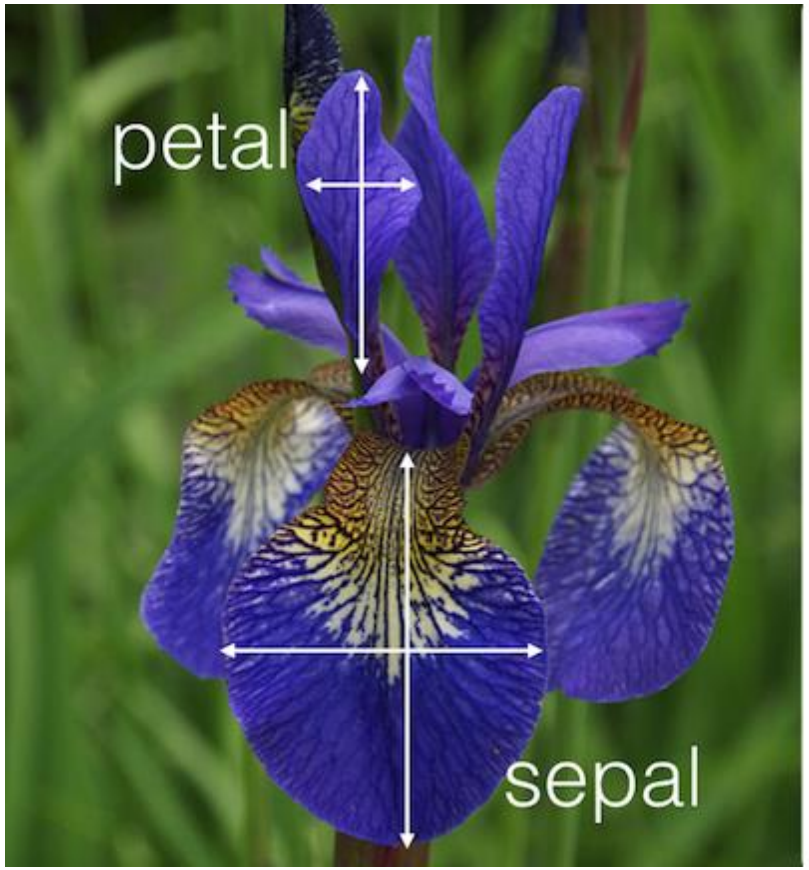


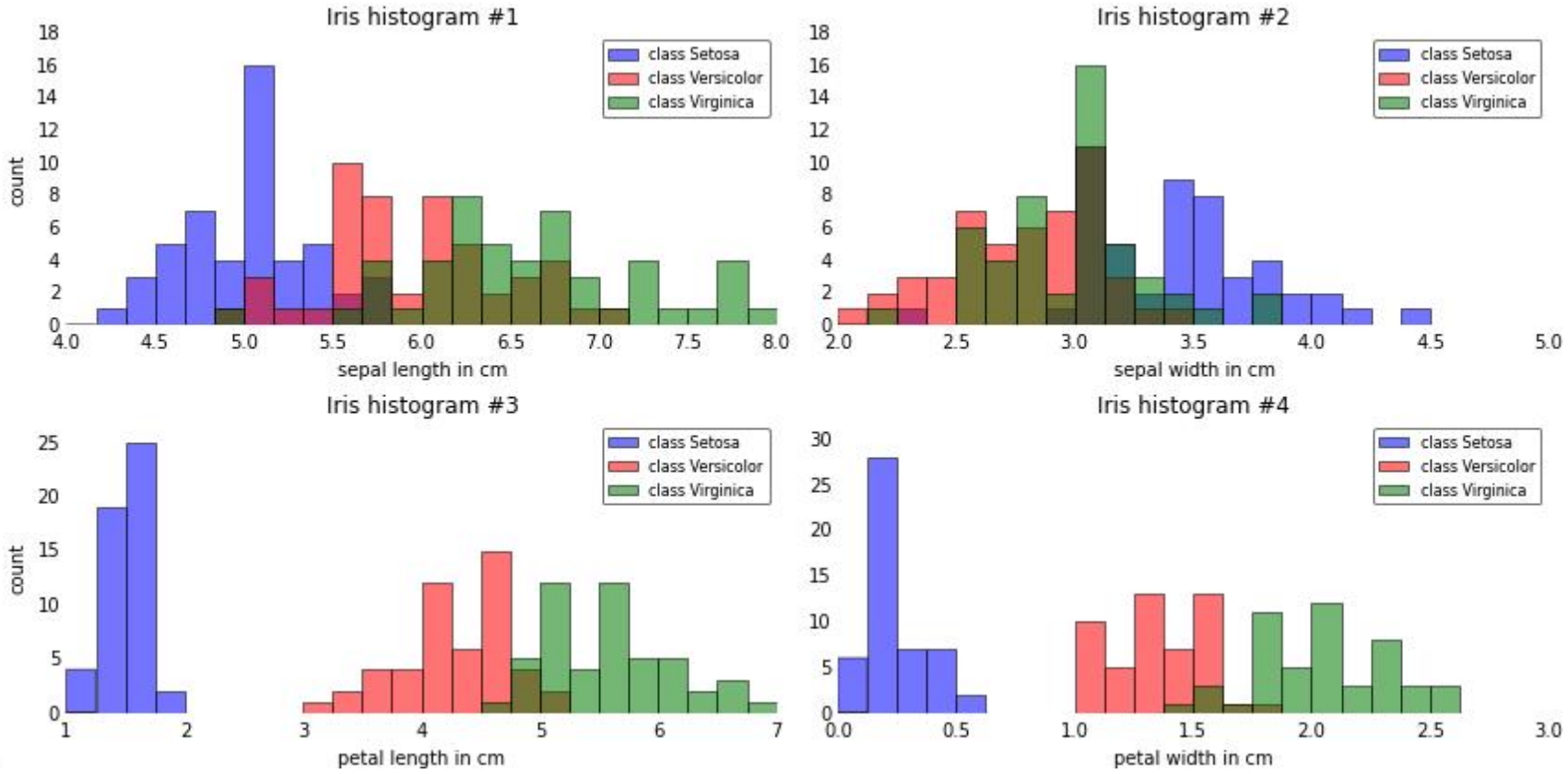
Linear regression

$$\hat{y} = \boldsymbol{x}\boldsymbol{\theta}$$



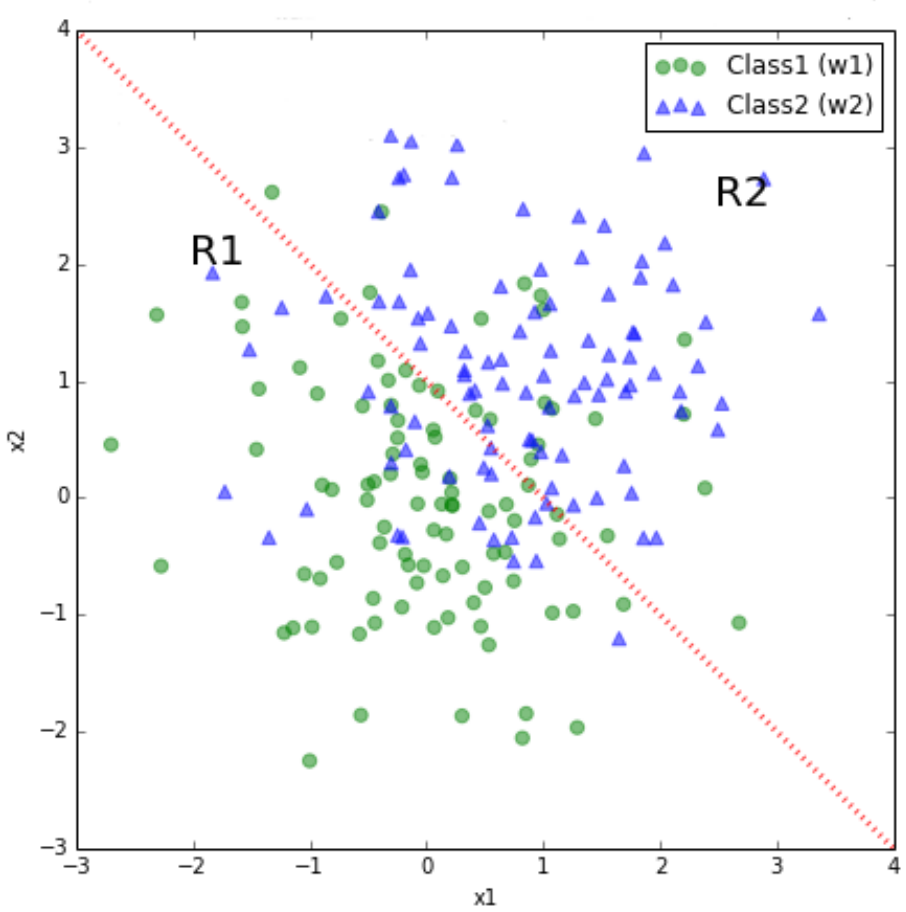
The Importance of Feature Selection

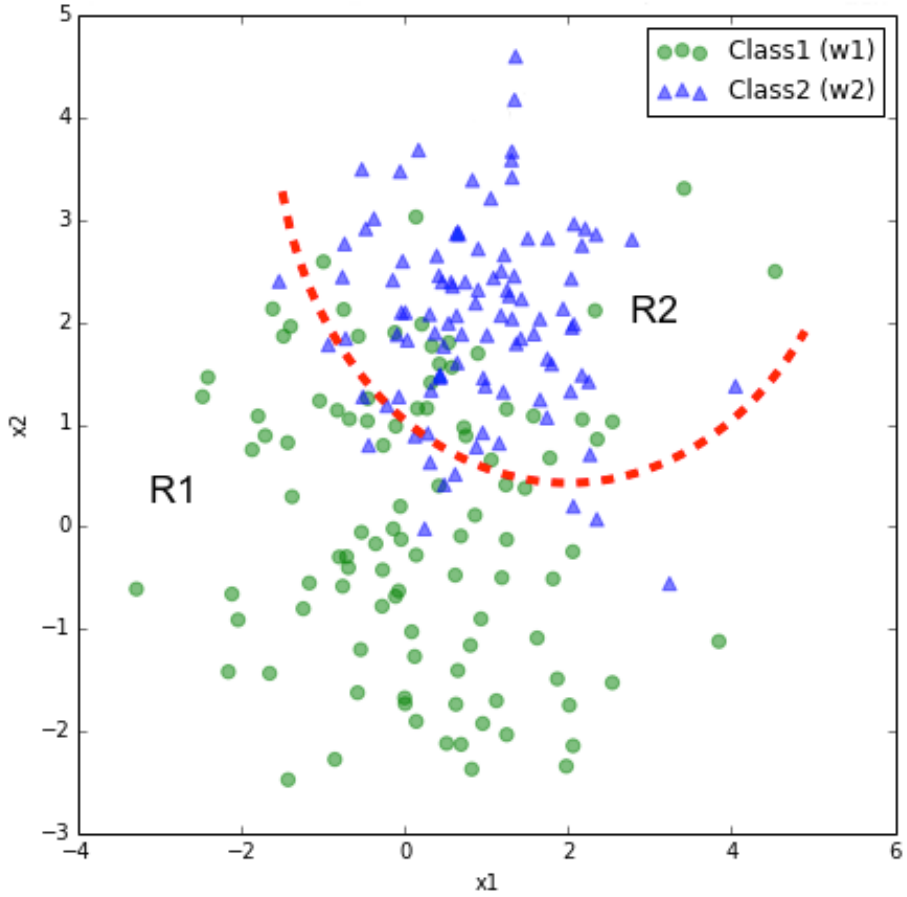




https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Classification

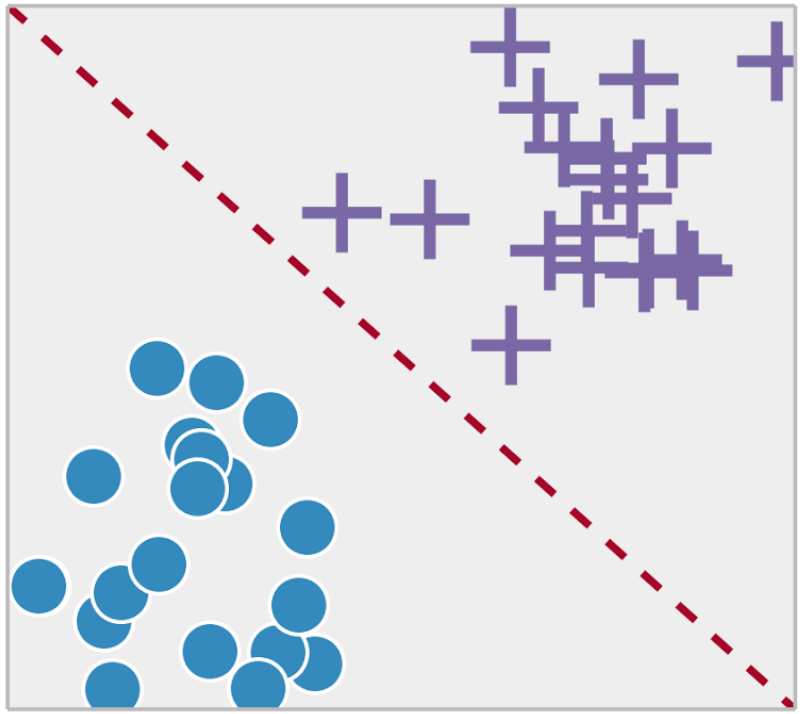




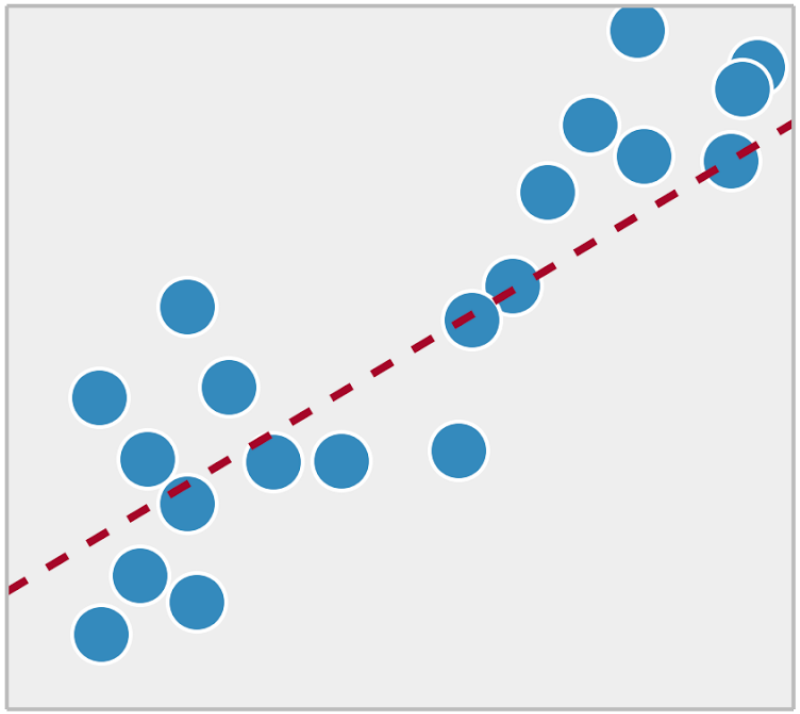
https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Model Types

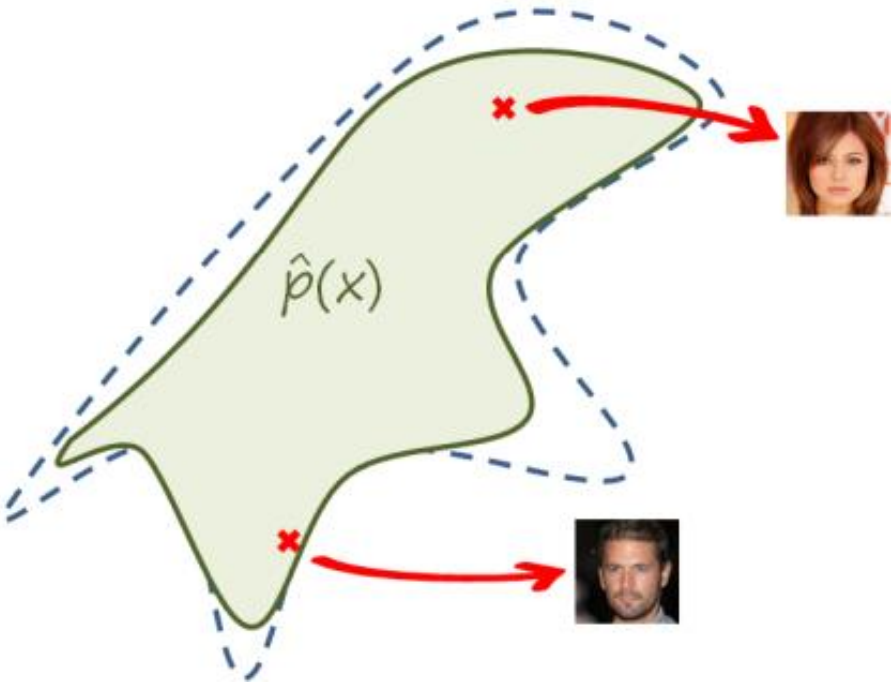
Classification



Regression

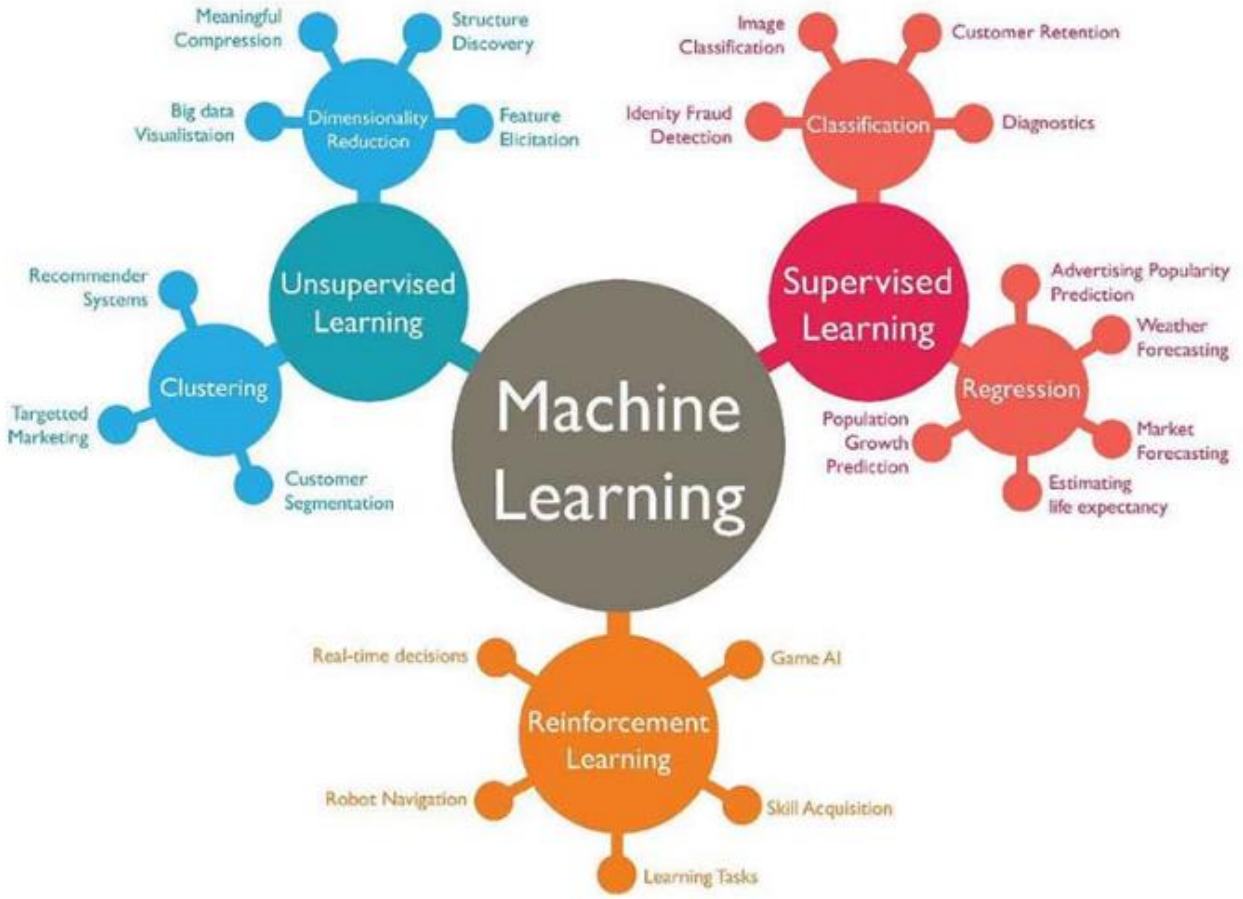


Generative



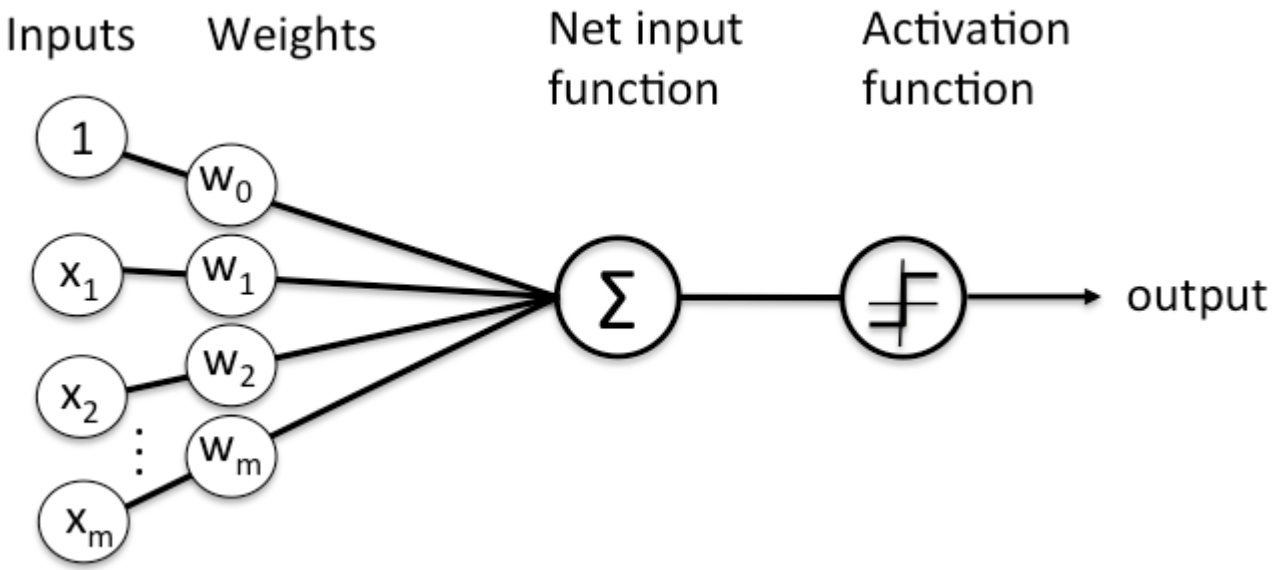
Types of Machine Learning

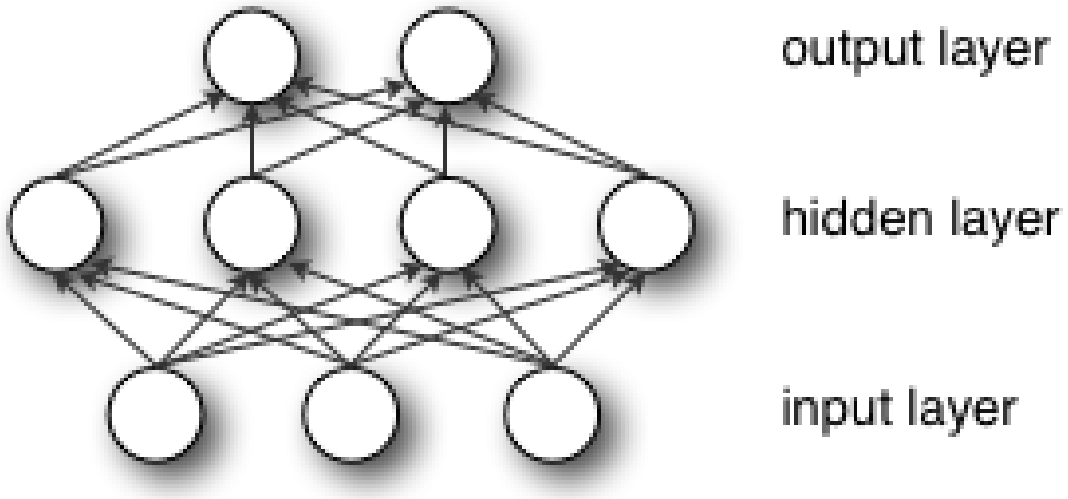
- Supervised learning
  - Given input-output examples f(X)=Y, learn the function f().
- Unsupervised learning
  - Given input examples, find patterns such as clusters
- Reinforcement learning
  - Select and execute an action, get feedback, update policy (what action to do in which state).



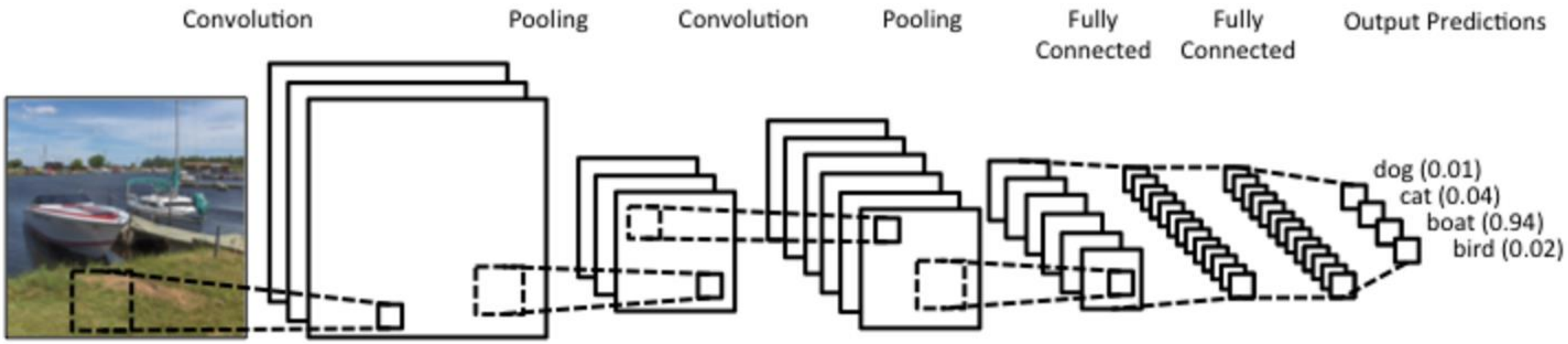
https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/

Neural Networks



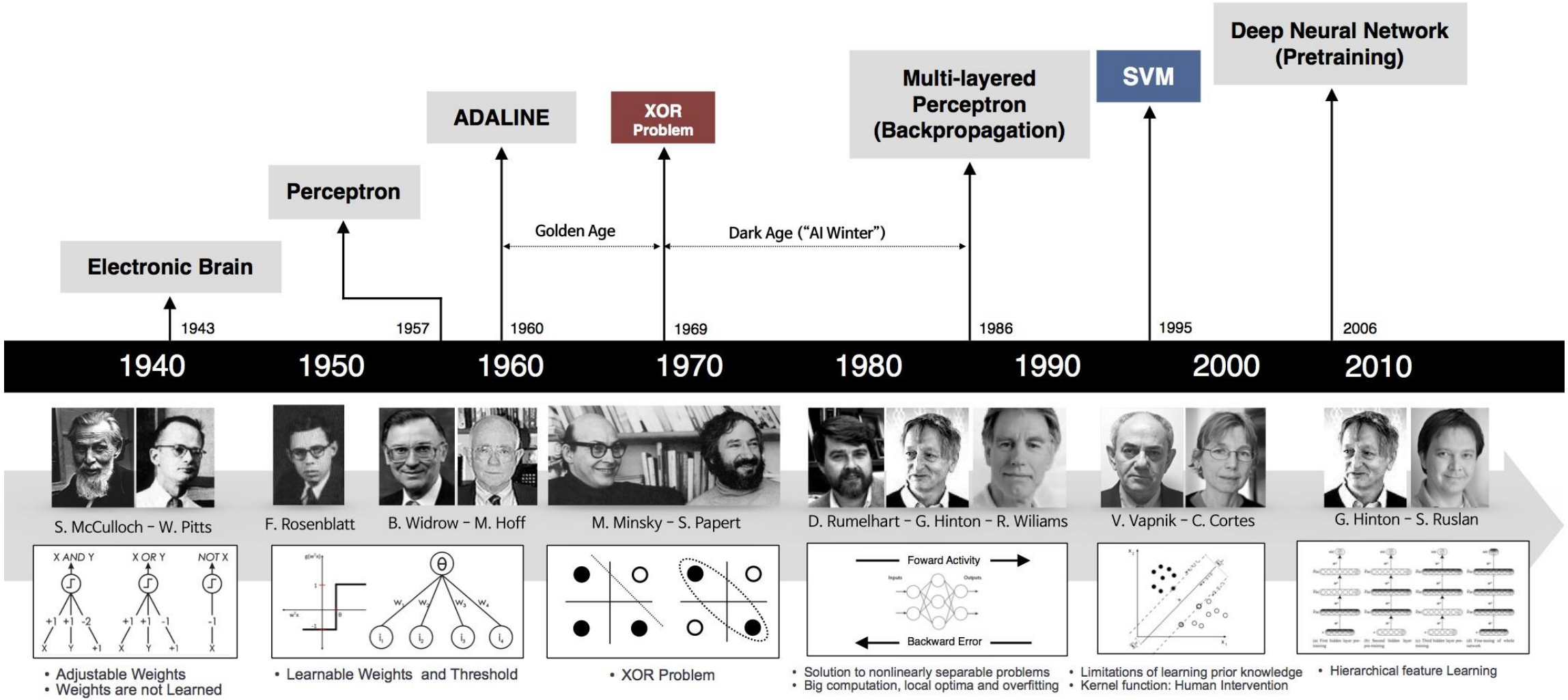


Deep Neural Networks



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

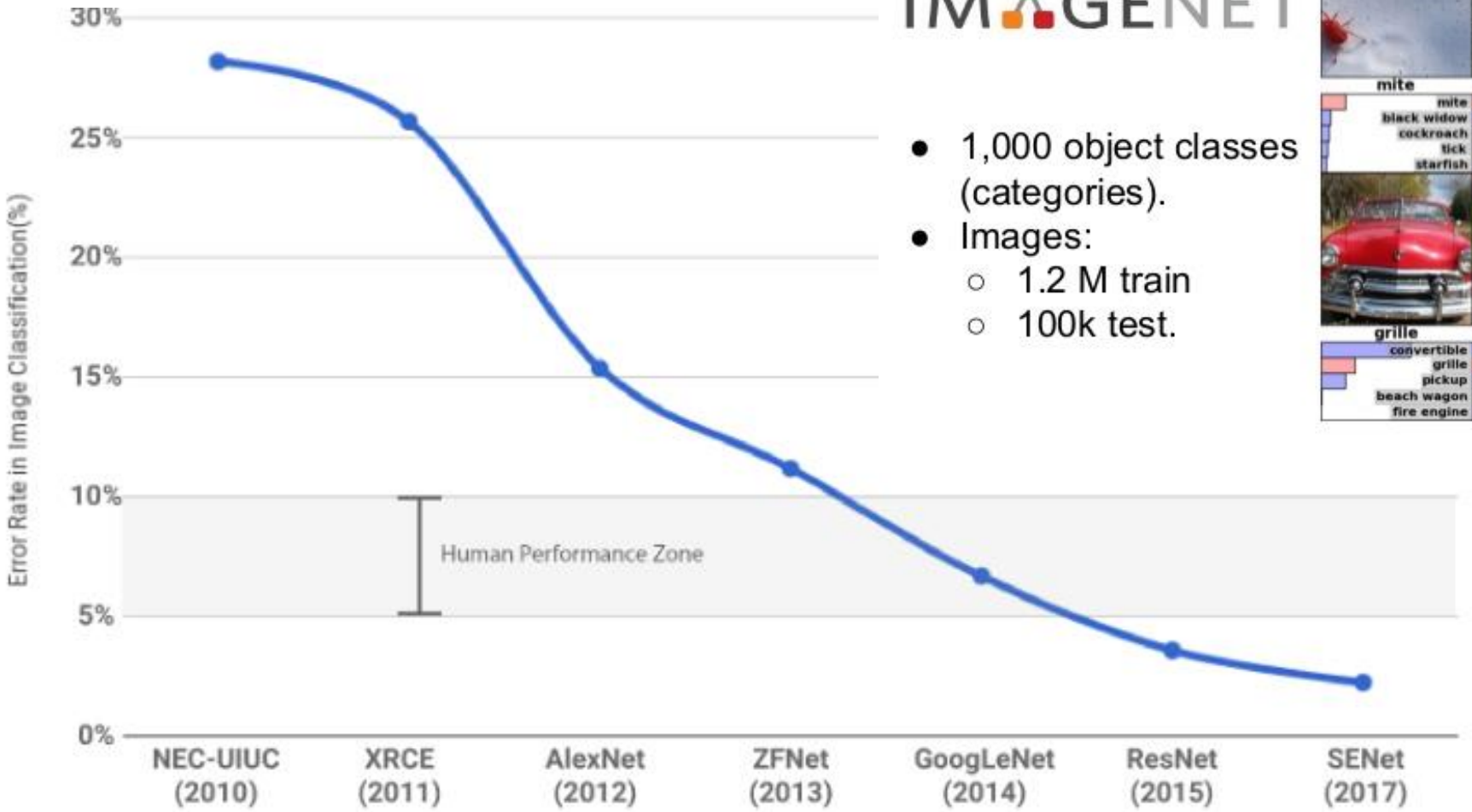
Neural Networks Timeline



LINKÖPINGS UNIVERSITET

https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

Image Classification

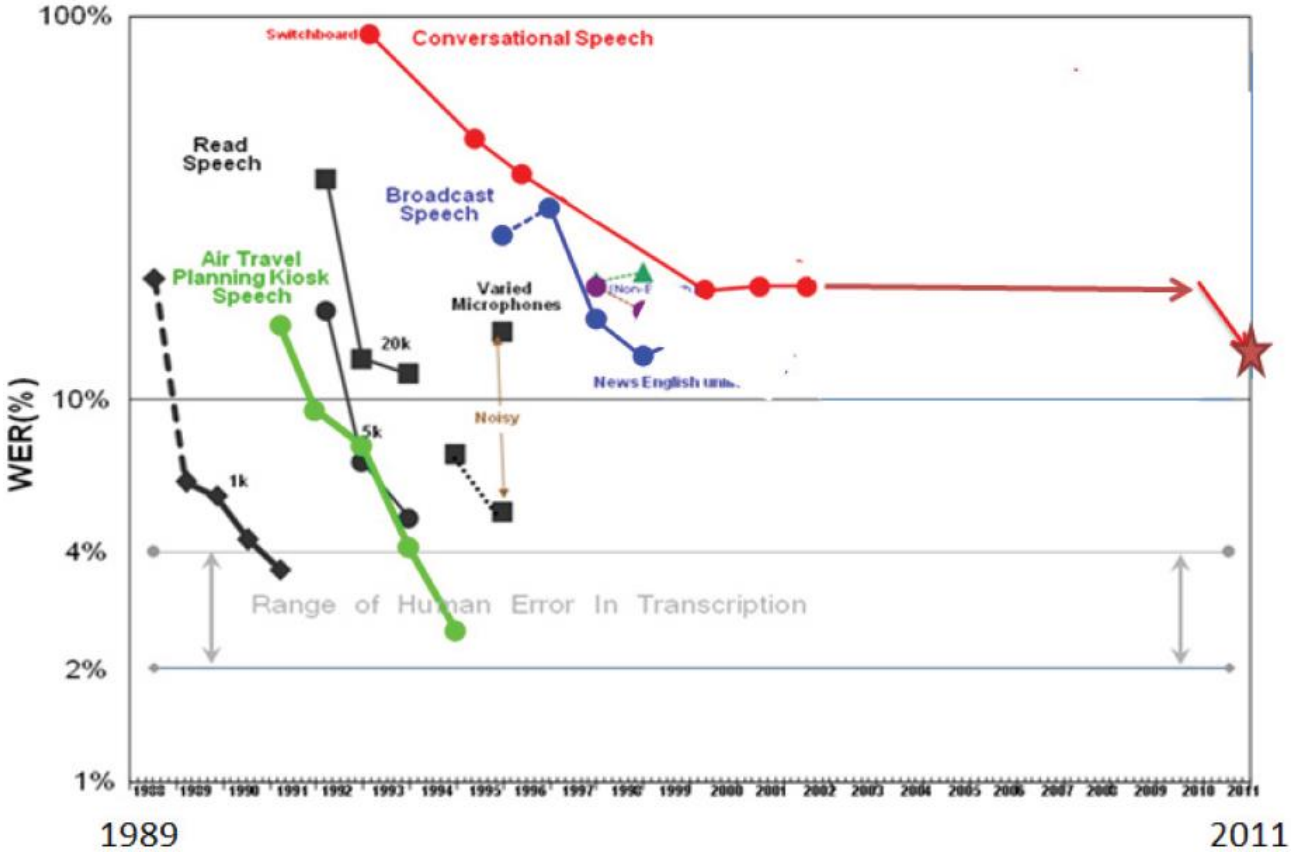


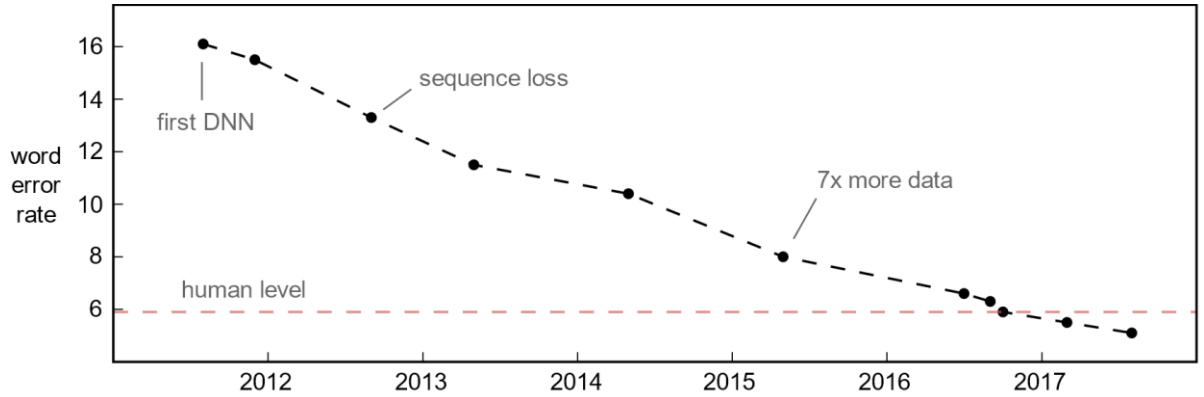
ImageNet Challenge

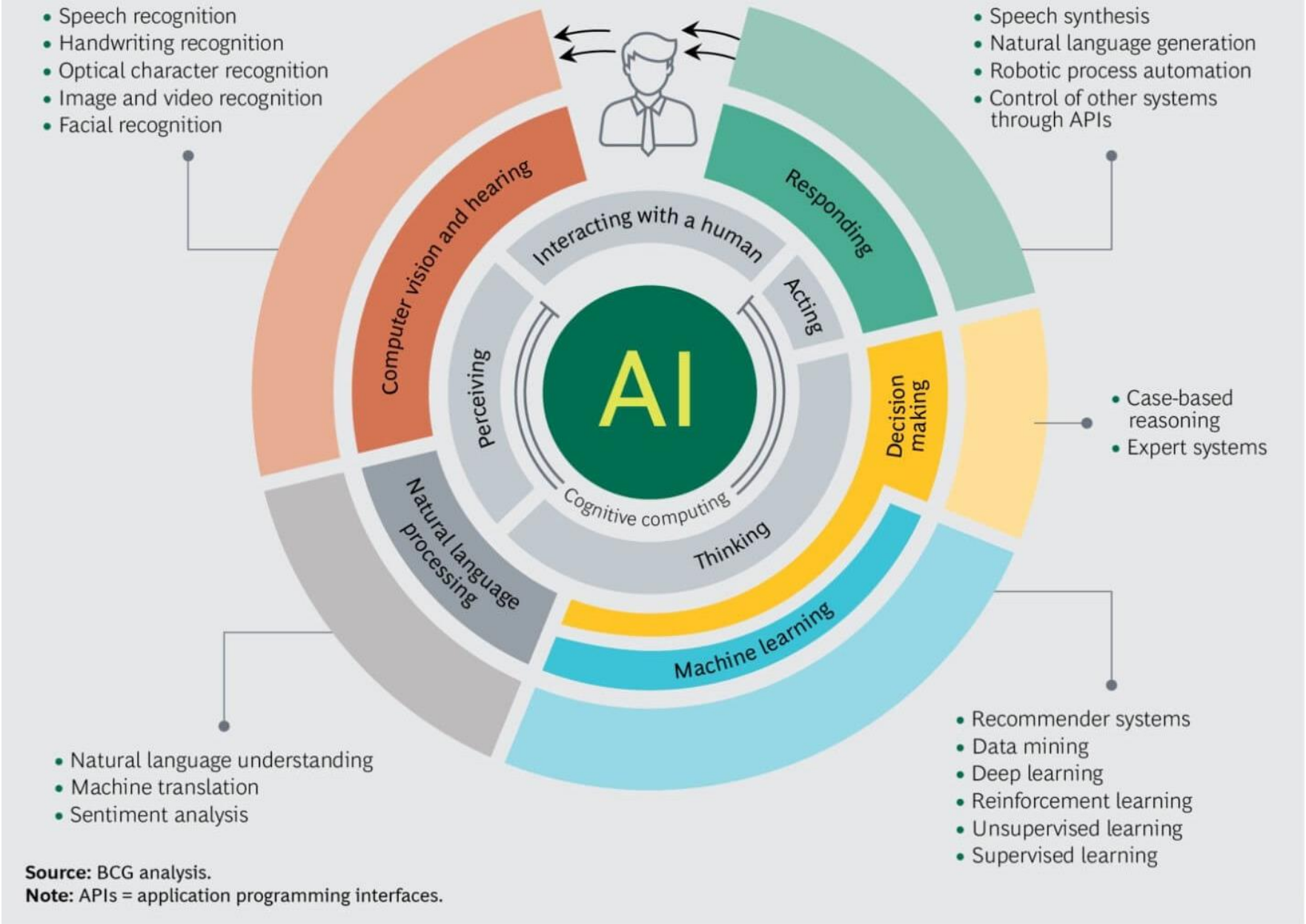
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

Speech Recognition

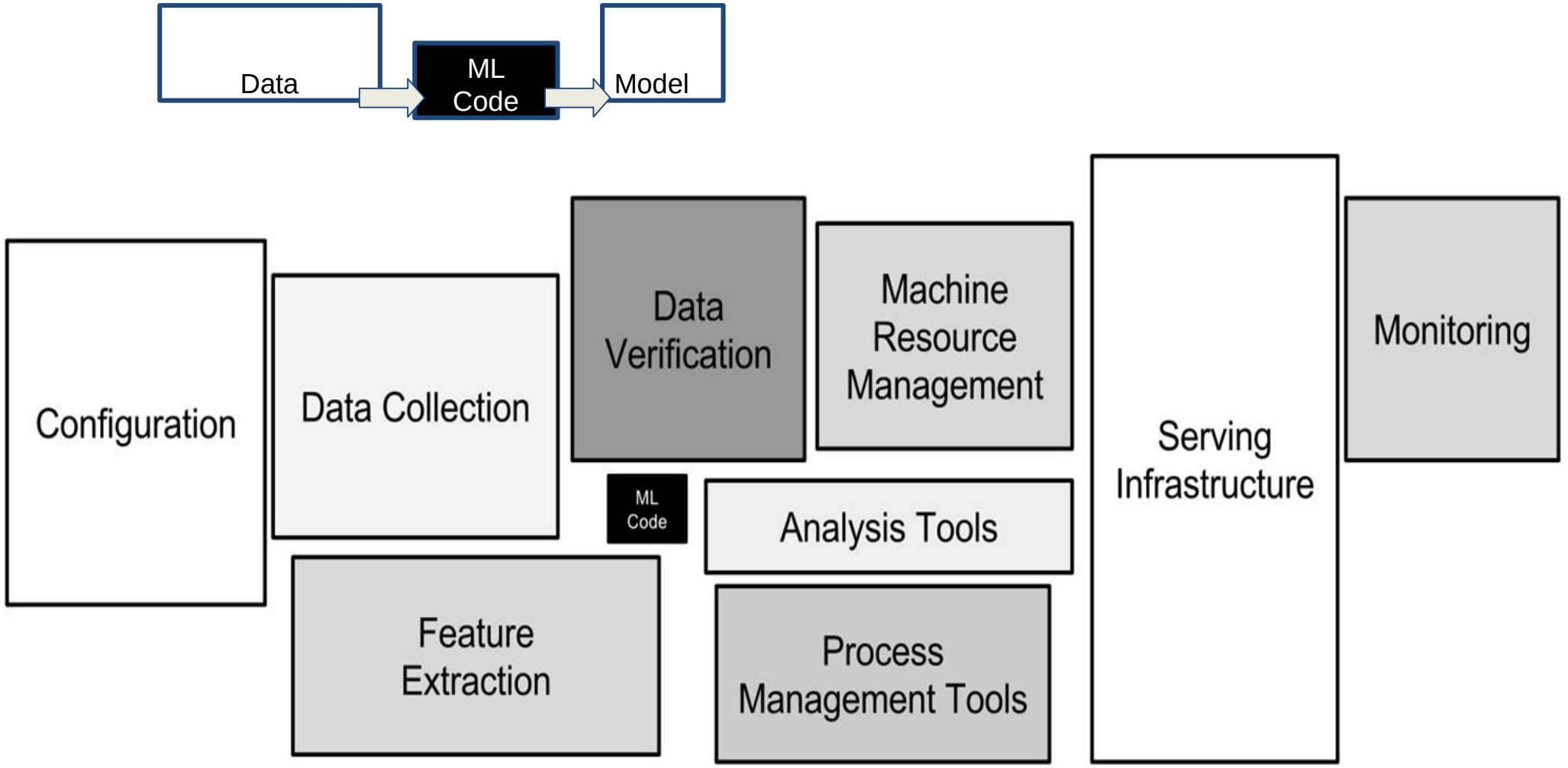




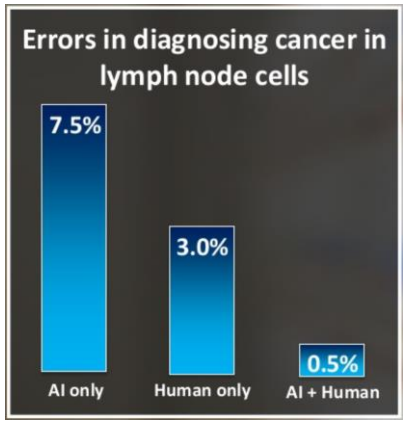


Source: BCG analysis.
Note: APIs = application programming interfaces.

The bigger system / picture



Hidden technical debt in Machine Learning Systems, Sculley et. al. (NIPS 2015)

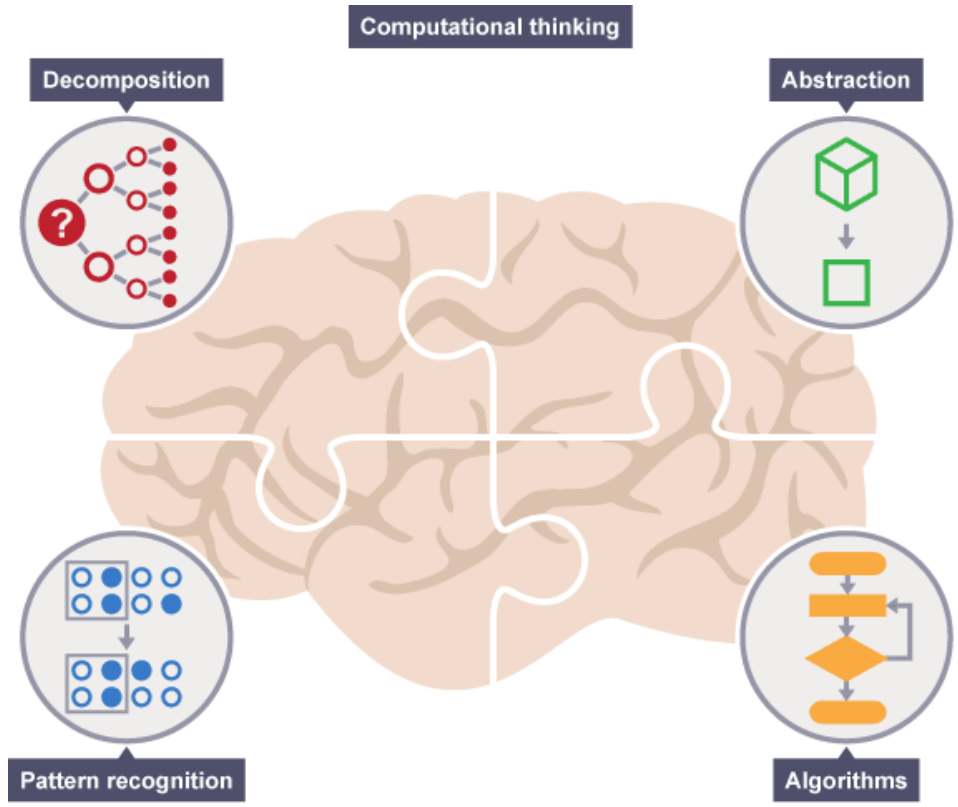


"Weak human + machine + superior process was greater than a strong computer and, remarkably, greater than a strong human + machine with inferior process."

Garry Kasparov

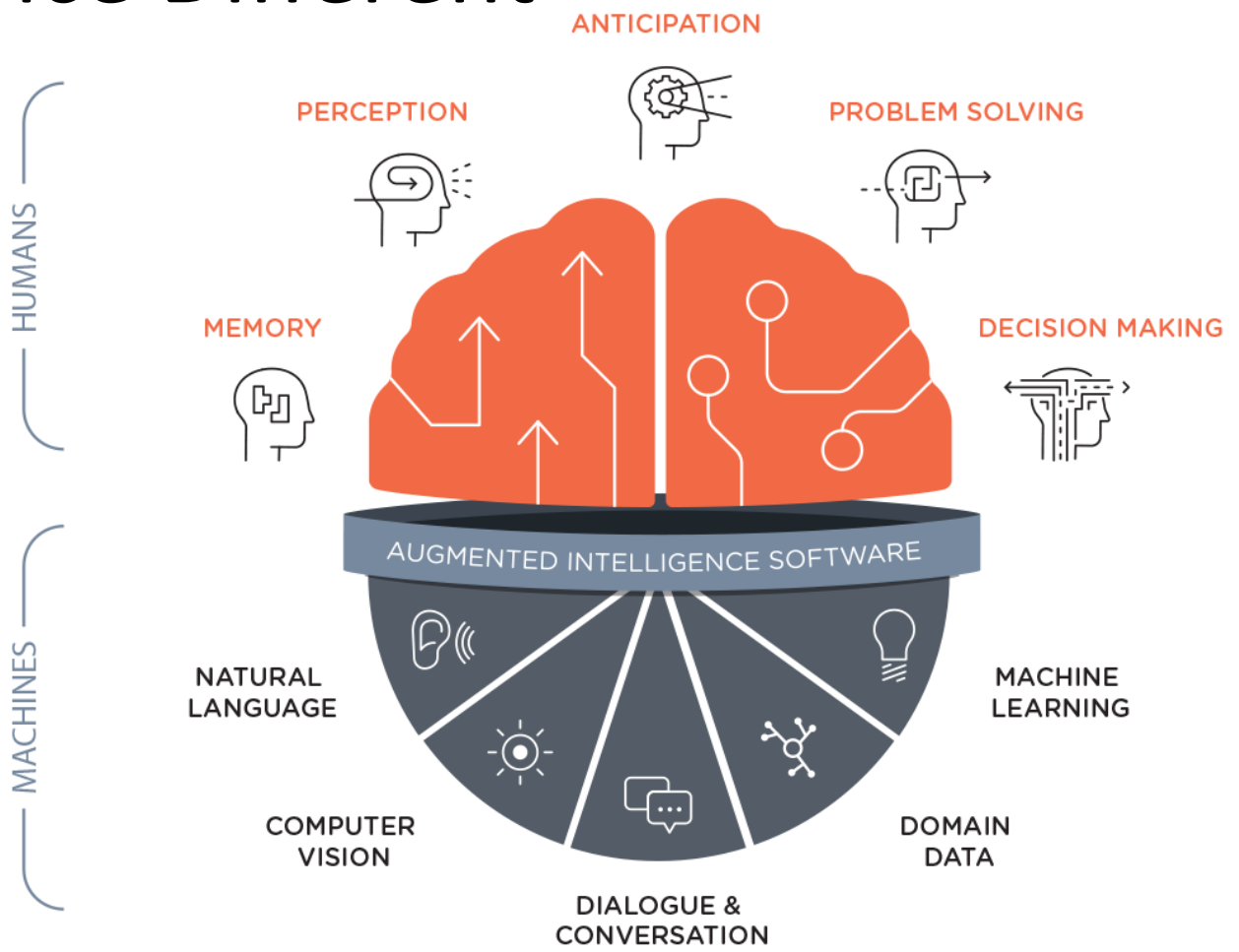
Computational Thinking

- Datalogiskt tänkande
- A **problem solving process** to describe, analyze, and solve problems such that computers can assist using techniques from computer science:
  - Give step-by-step instructions
  - Decompose problems into smaller parts
  - Find patterns
  - Create abstractions
  - Design algorithms



Why is Artificial Intelligence Different

- Scale
- Speed
- Single-mindedness
- Optimization-based
- Cannot break the rules
- No needs
- No real consequences or “skin in the game”



EU strategy for AI

A STRATEGY FOR EUROPE TO LEAD THE WAY

Boost technological and industrial capacity & AI uptake

Prepare for socio-economic changes

Ensure an appropriate ethical & legal framework

In this context: appointment of Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018

Ethics Guidelines for Trustworthy AI – Overview

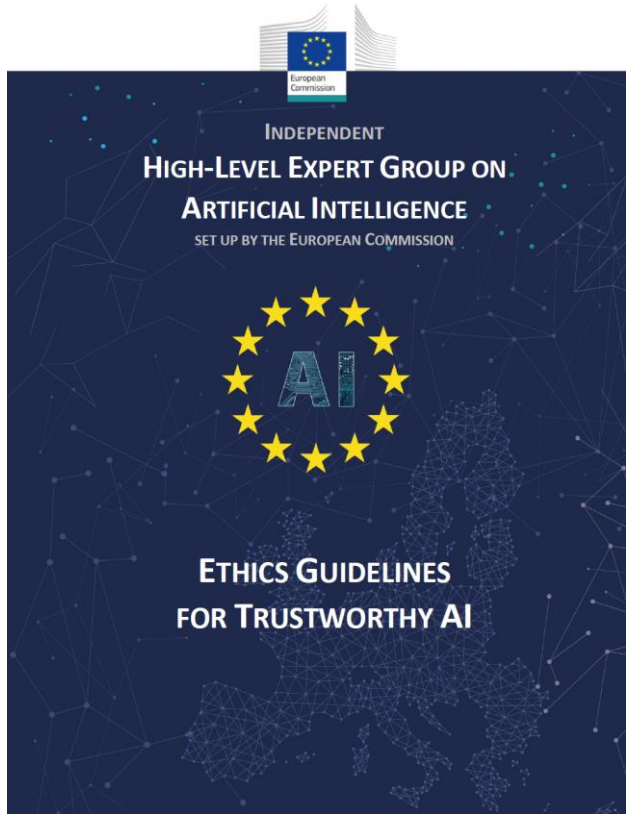
Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

- Lawful AI
- Ethical AI
- Robust AI

Three levels of abstraction

- from principles (Chapter I)
- to requirements (Chapter II)
- to assessment list (Chapter III)



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ethics Guidelines for Trustworthy AI – Principles

4 Ethical Principles based on fundamental rights

Respect for human autonomy

Augment, complement and empower humans

Prevention of harm

Safe and secure. Protect physical and mental integrity.

Fairness

Equal and just distribution of benefits and costs.

Explicability

Transparent, open with capabilities and purposes, explanations

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ethics Guidelines for Trustworthy AI – Requirements

Human agency and oversight

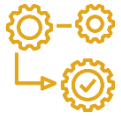
Technical Robustness and safety

Privacy and data governance

Transparency

Diversity, non-discrimination and fairness

Societal & environmental well-being



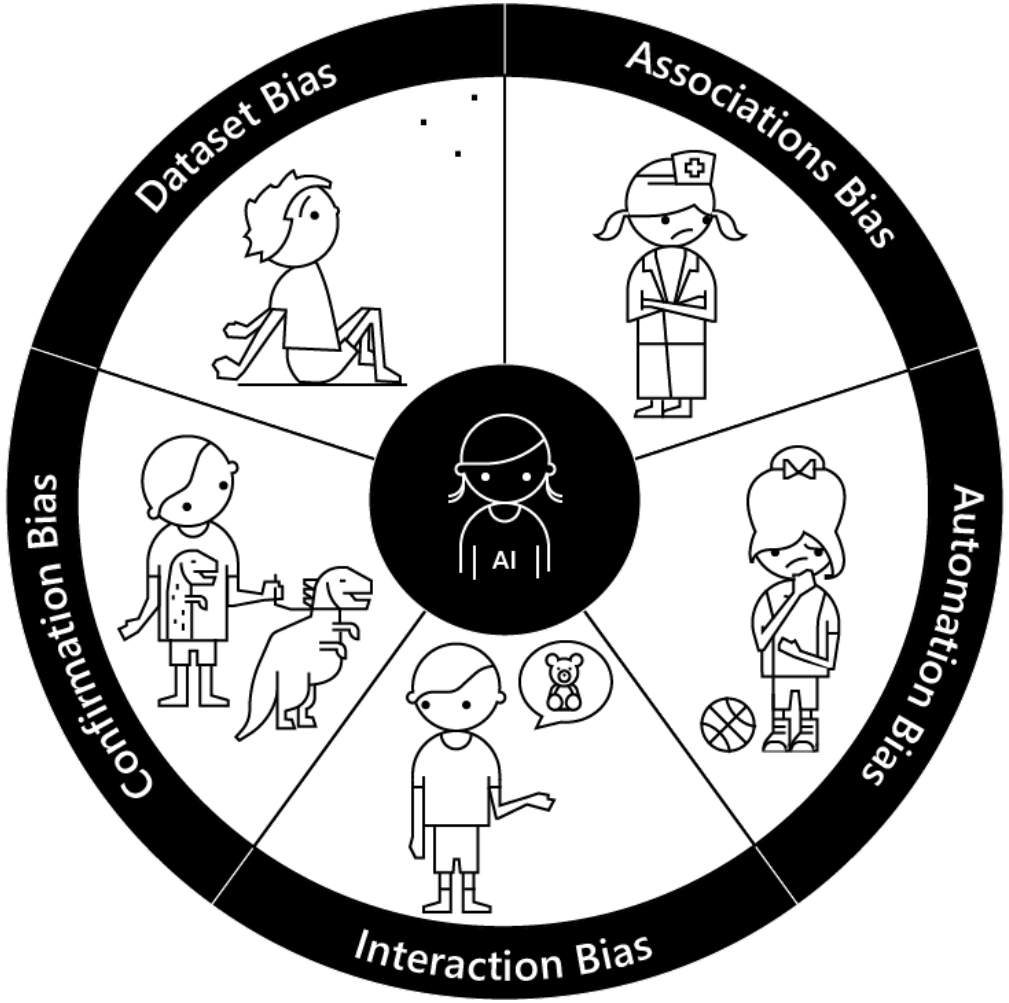
Accountability

To be continuously implemented & evaluated throughout AI system's life cycle

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Bias

- **Dataset bias** – When the data used to train machine learning models doesn't represent the diversity of the customer base.
- **Association bias** – When the data used to train a model reinforces and multiplies a cultural bias.
- **Automation bias** – When automated decisions override social and cultural considerations.
- **Interaction bias** – When humans tamper with AI and create biased results.
- **Confirmation bias** – When oversimplified personalization makes biased assumptions for a group or an individual.



https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850

Research Challenges

Respect for human autonomy

- Human-AI interaction
- Meaningful human control

Fairness

- Fairness

Prevention of harm

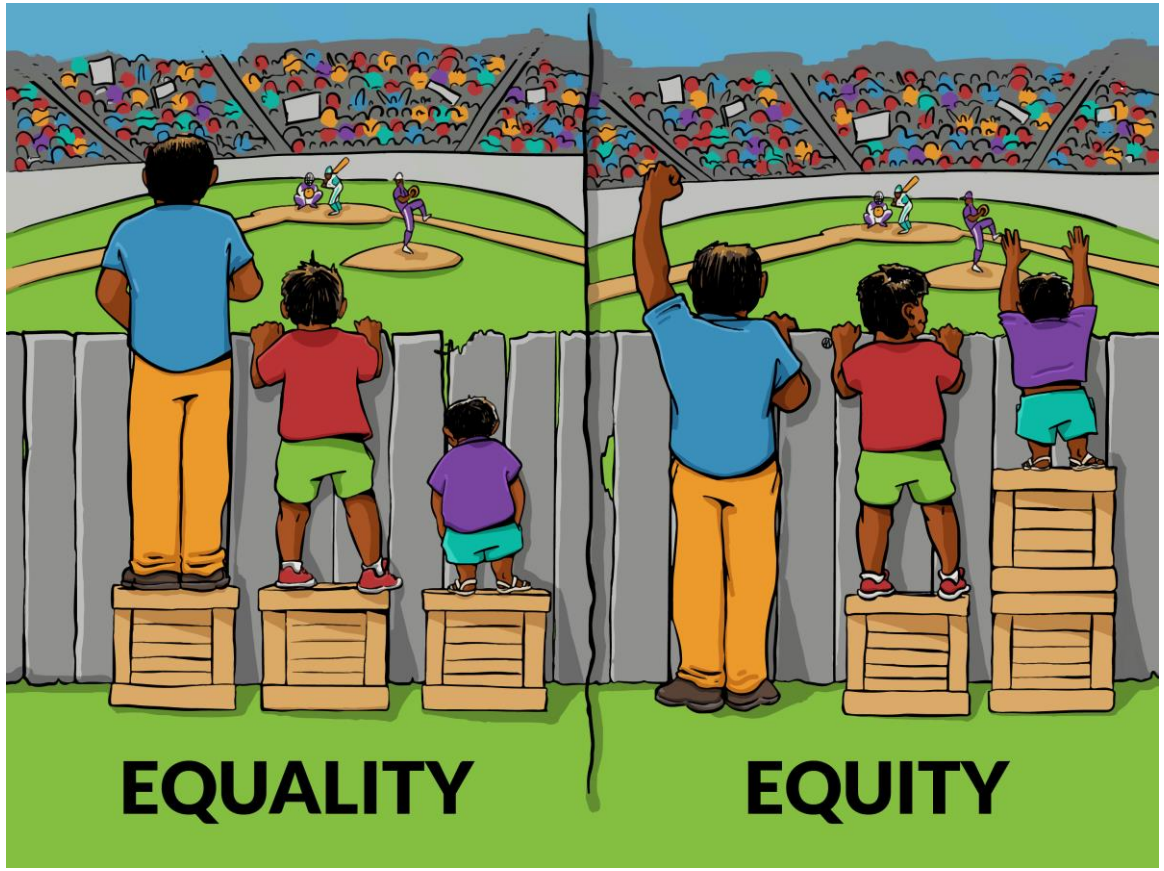
- AI Safety
- Value alignment

Explicability

- Accountability
- Transparency
- Explainability

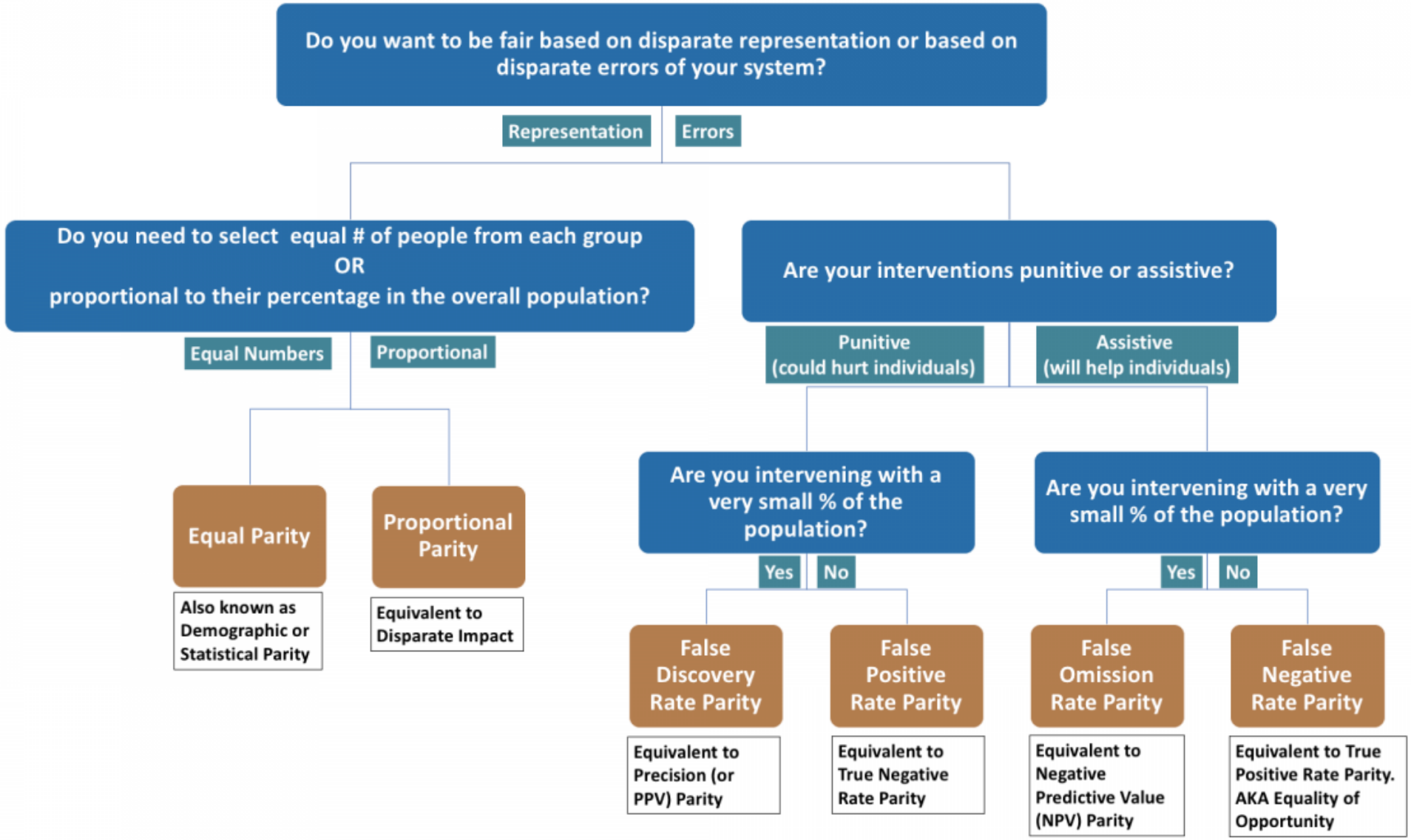
What is Fairness and Fairness-Aware ML?

- Impartial and just treatment or behavior without favoritism or discrimination.
- *Fairness-aware machine learning* algorithms seek to provide methods under which the predicted outcome of a classifier operating on data about people is fair or non-discriminatory.

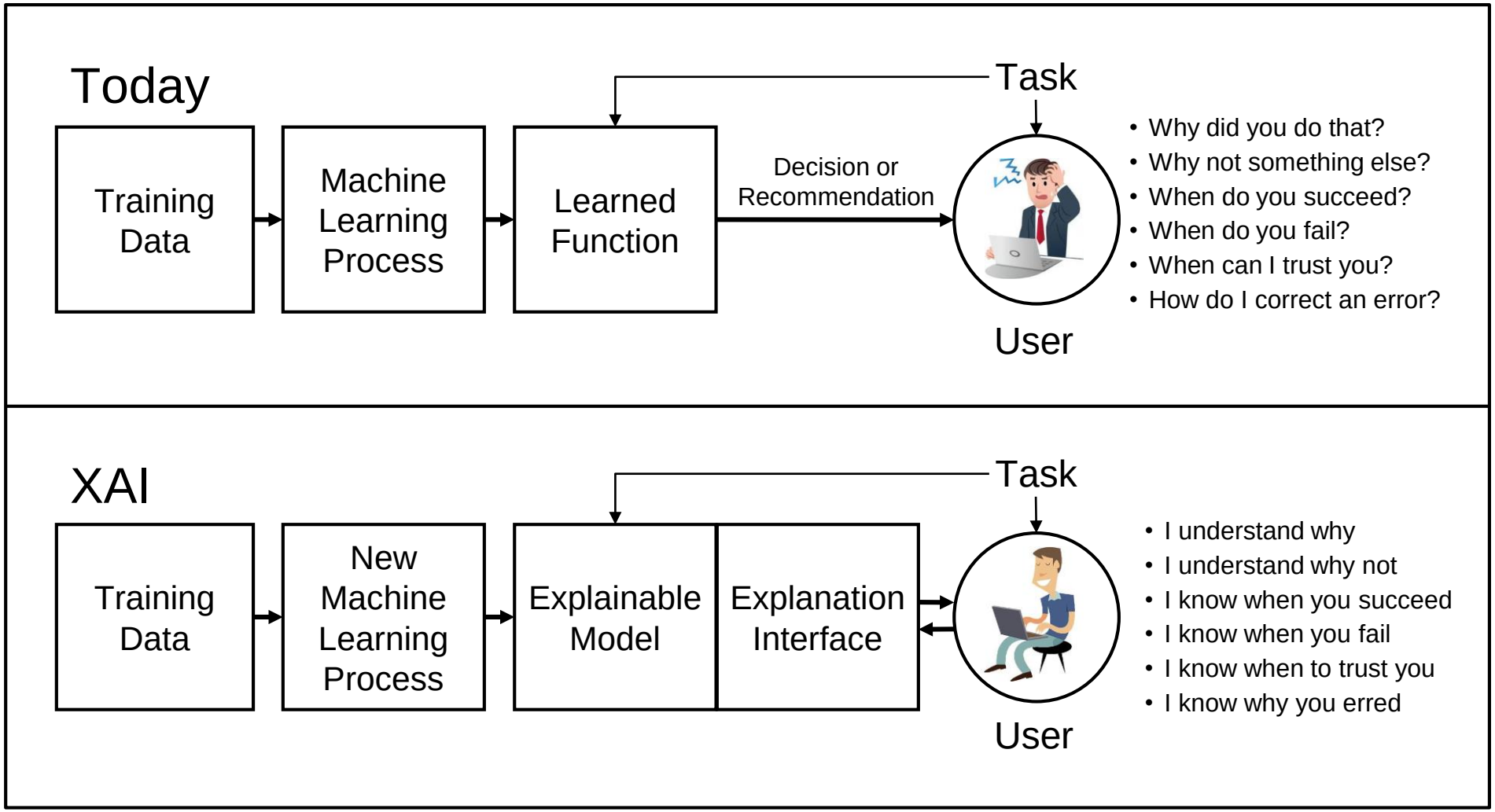


Interaction Institute for Social Change | Artist: Angus Maguire

FAIRNESS TREE



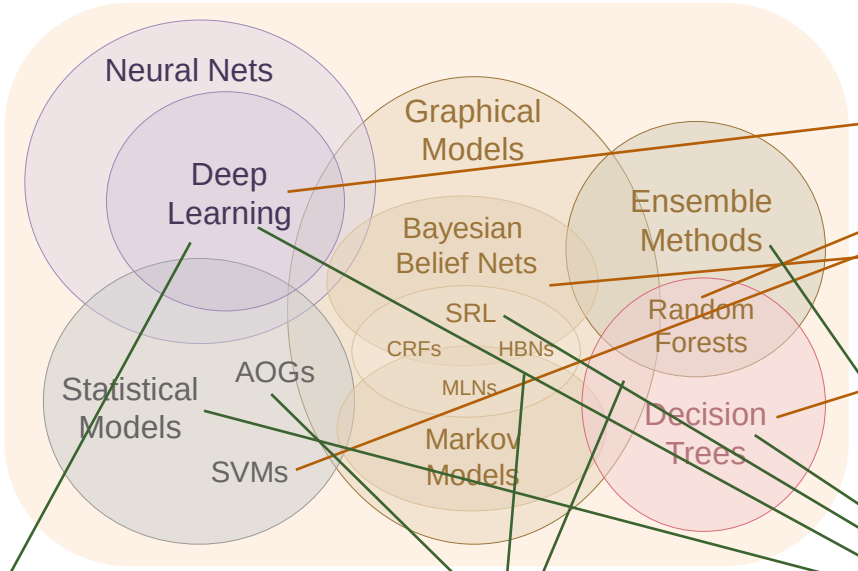
Explainable AI – The DARPA View



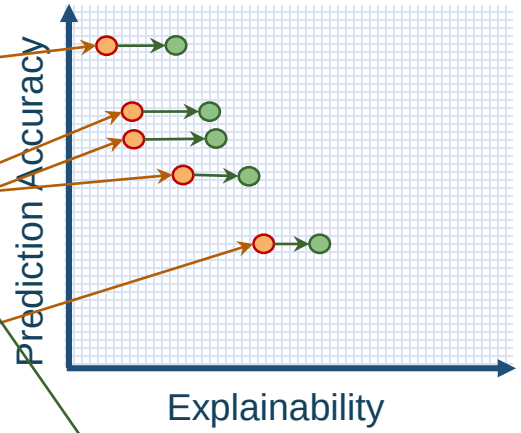
New Approach

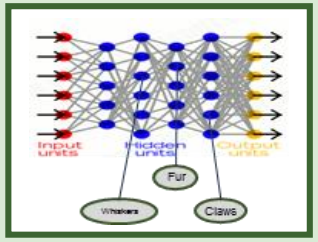
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



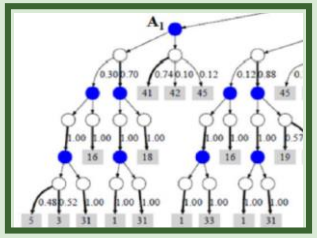
Explainability (notional)





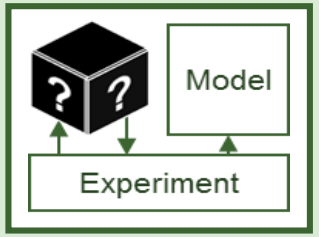
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

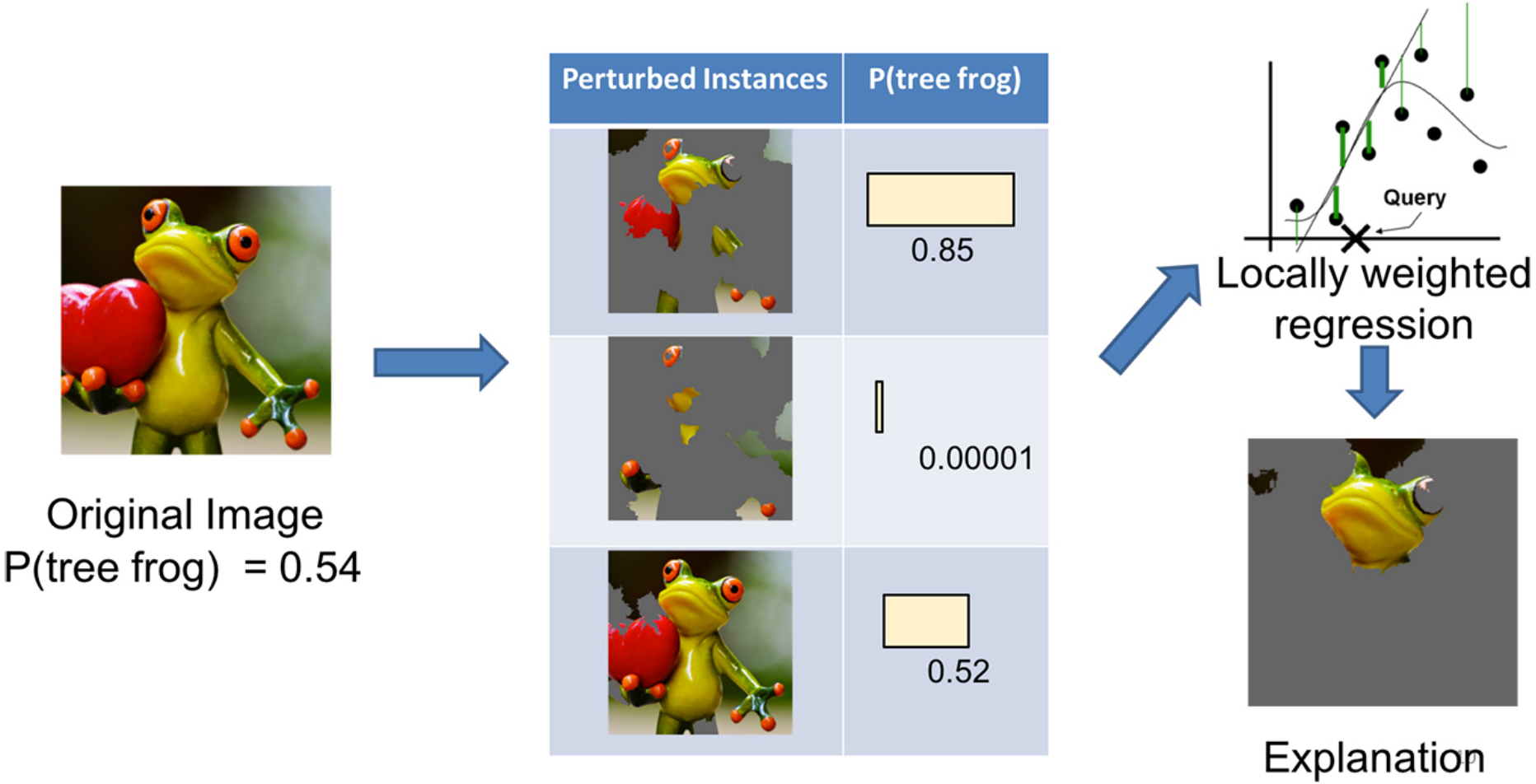
Techniques to learn more structured, interpretable, causal models



Model Induction

Techniques to infer an explainable model from any model as a black box

LIME (Local Interpretable Model-agnostic Explanations)



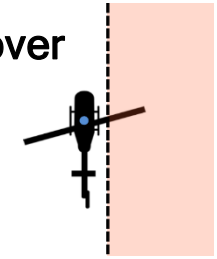
https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Safe Autonomous Systems / AI

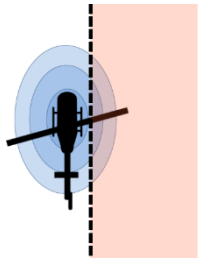
If things can go wrong they probably will!

This implies the need for continual monitoring of an autonomous system and its environment in a principled, contextual, task specific manner which can be specified by the system itself!

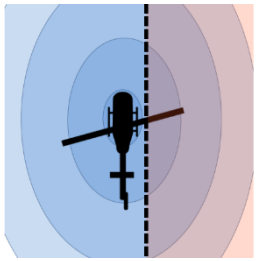
Reasoning over Uncertainty



collision: false

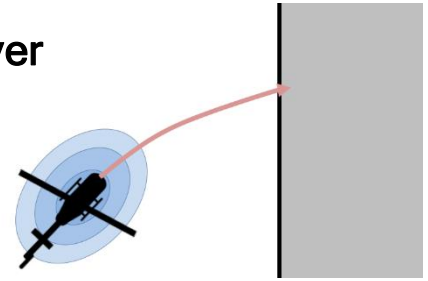


Pr(collision) = 0.1

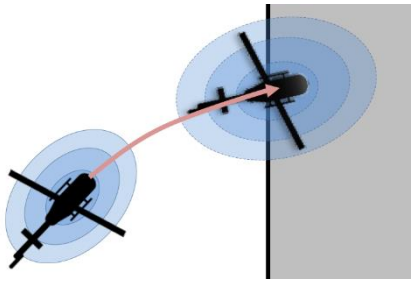


Pr(collision) = 0.4

Reasoning over Predictions



Pr(collision now) = 0.0...

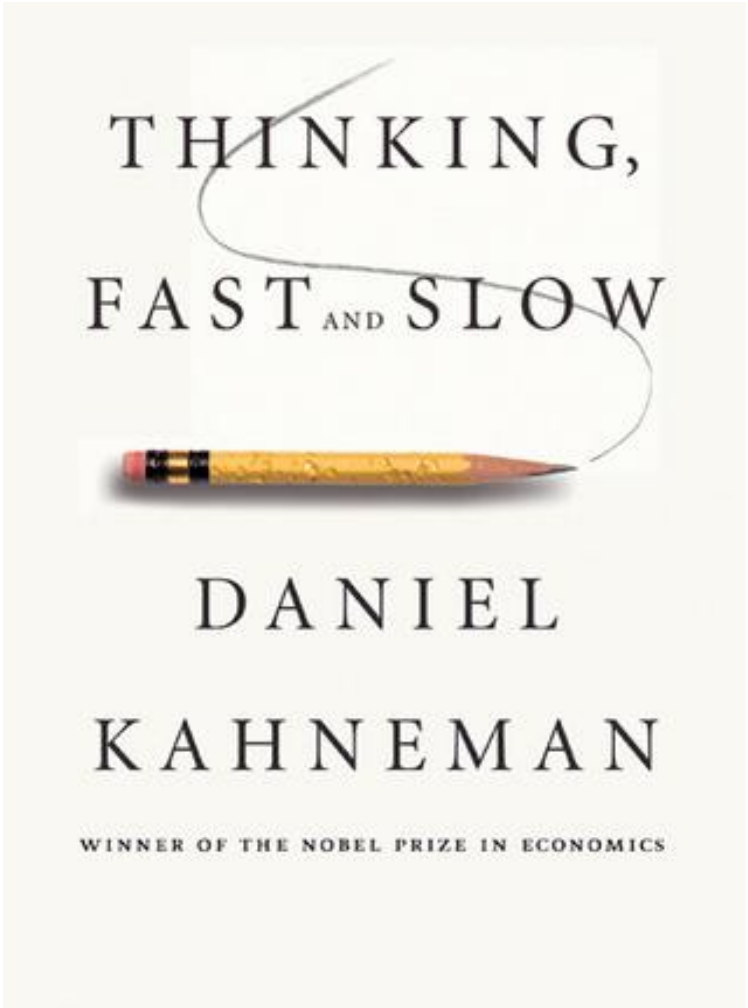


Pr(collision soon) = 0.5

Human and Computational Thinking

Figure 1: A Comparison of System 1 and System 2 Thinking

System 1 "Fast"	System 2 "Slow"
DEFINING CHARACTERISTICS Unconscious Effortless Automatic	DEFINING CHARACTERISTICS Deliberate and conscious Effortful Controlled mental process
WITHOUT self-awareness or control	WITH self-awareness or control
"What you see is all there is."	Logical and skeptical
ROLE Assesses the situation Delivers updates	ROLE Seeks new/missing information Makes decisions



Pure Logic **Pure Learning**

- Slow thinking: deliberative, cognitive, model-based, extrapolation
- Amazing achievements until this day
- "*Pure logic is brittle*"
  noise, uncertainty, incomplete knowledge, …

https://web.cs.ucla.edu/~guyvdb/slides/ComputersAndThought.pdf

Pure Logic **Pure Learning**

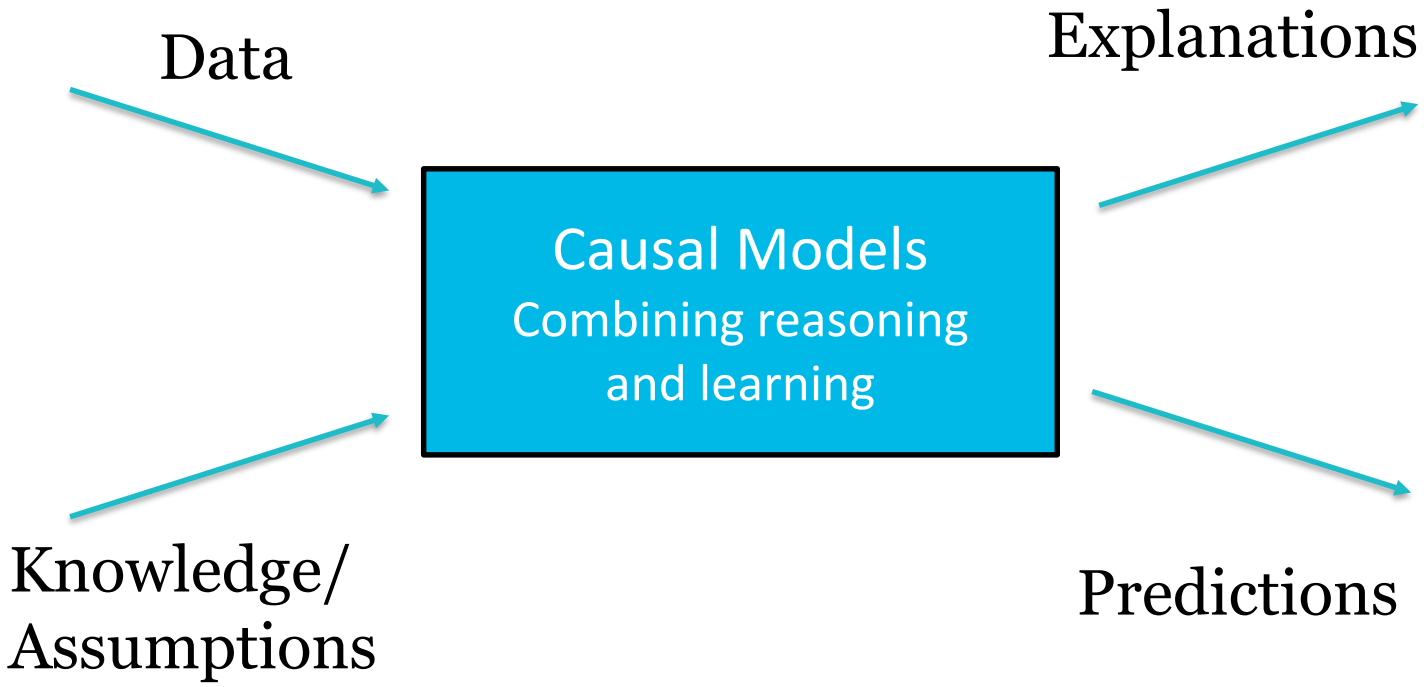
- Fast thinking: instinctive, perceptive, model-free, interpolation
- Amazing achievements recently
- "*Pure learning is brittle*"

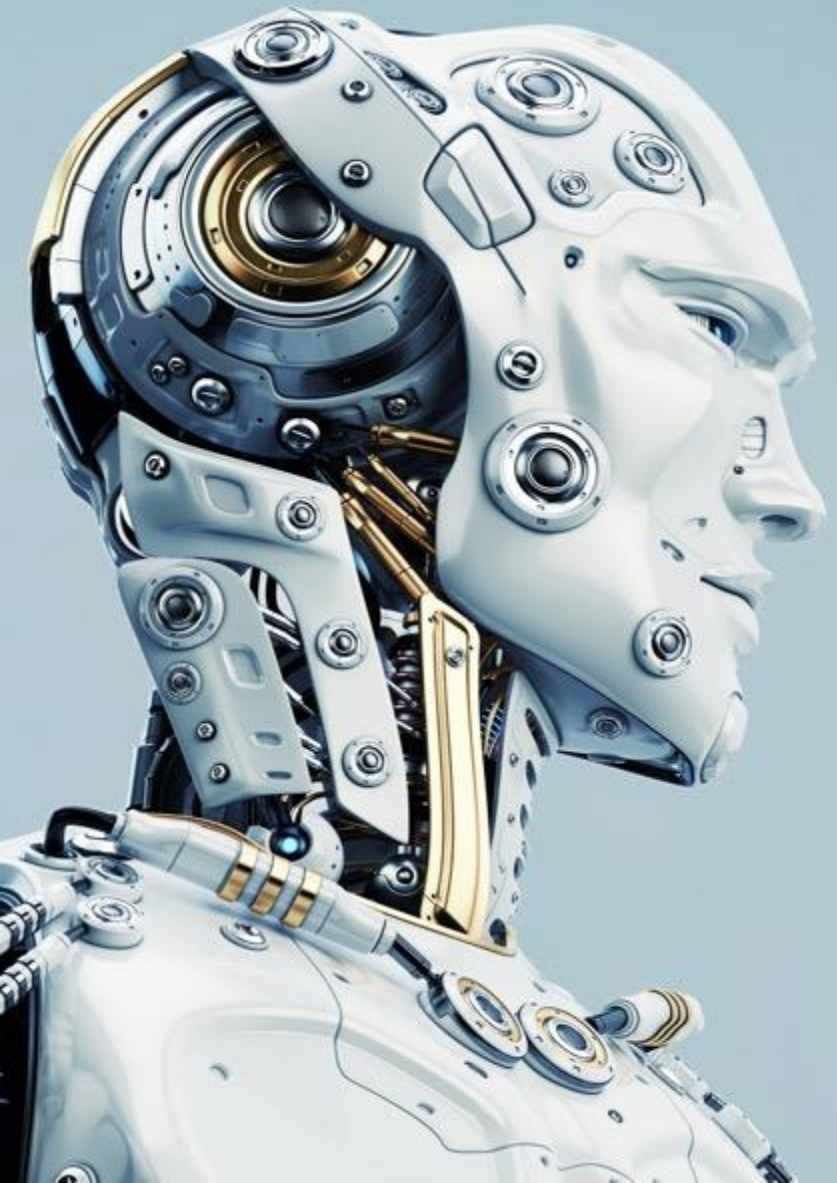
  bias, algorithmic fairness, interpretability, explainability, adversarial attacks, unknown unknowns, calibration, verification, missing features, missing labels, data efficiency, shift in distribution, general robustness and safety

  fails to incorporate a sensible model of the world

https://web.cs.ucla.edu/~guyvdb/slides/ComputersAndThought.pdf

The Way Forward

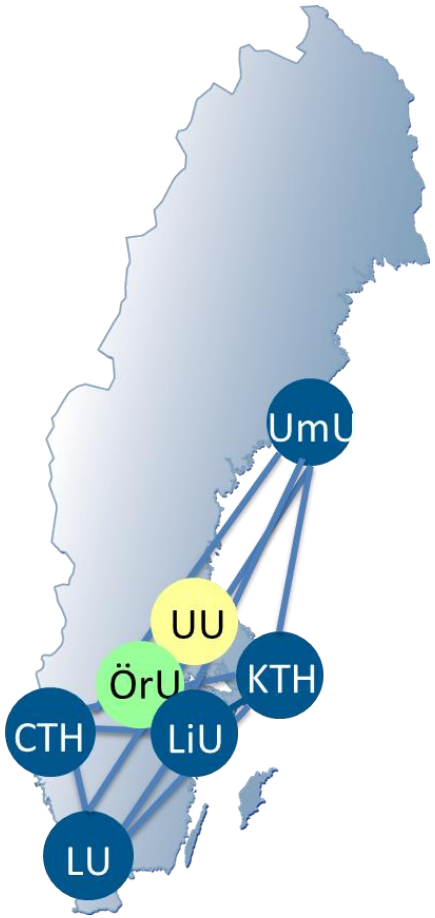




"AI systems should help empower society, combining the best of technology with the best of humanity"

http://www.aisustainability.org/

AI Innovation, Competence and Research in Sweden



AI SUSTAINABILITY CENTER

AI INNOVATION of Sweden

AI Competence of Sweden

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

CHALMERS UNIVERSITY OF TECHNOLOGY · KTH VETENSKAP OCH KONST · LINKÖPINGS UNIVERSITET · LUNDS UNIVERSITET · UMEÅ UNIVERSITET · UPPSALA UNIVERSITET · ÖREBRO UNIVERSITY

RI.SE

Elements of AI

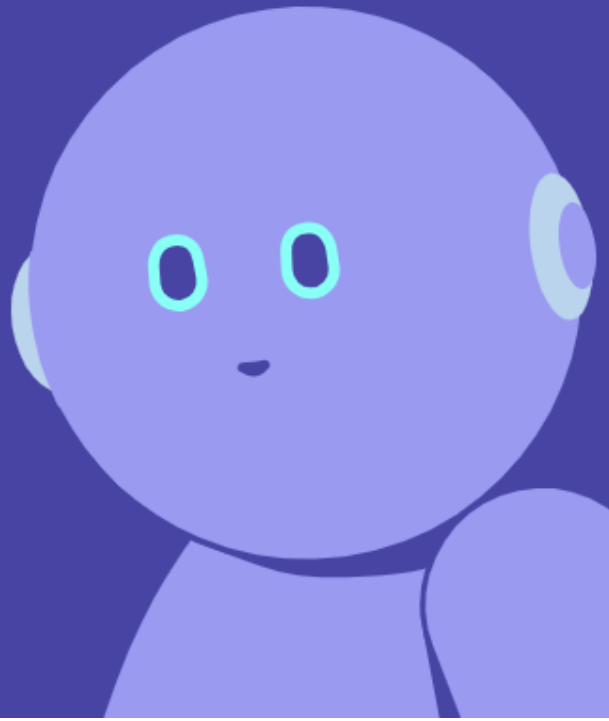
Elements of AI

Finnish →

Welcome to the Elements of Artificial Intelligence free online course

English ▾ | Start the course

Distance course at Linköping University to get 2ECTS



UNIVERSITY OF HELSINKI

Reaktor

LINKÖPINGS UNIVERSITET

AI INNOVATION of Sweden

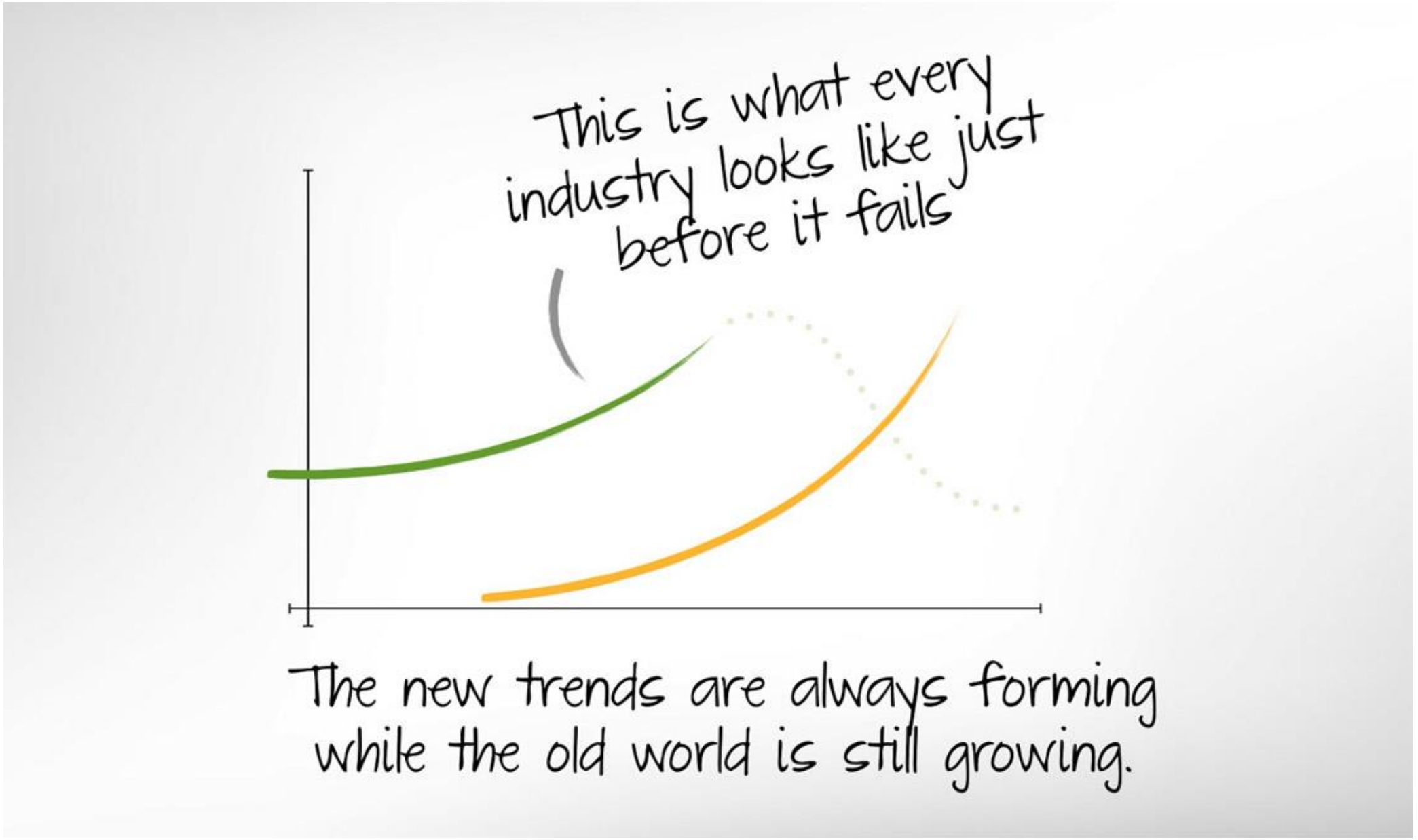
AI COMPETENCE FOR SWEDEN

https://www.elementsofai.se/

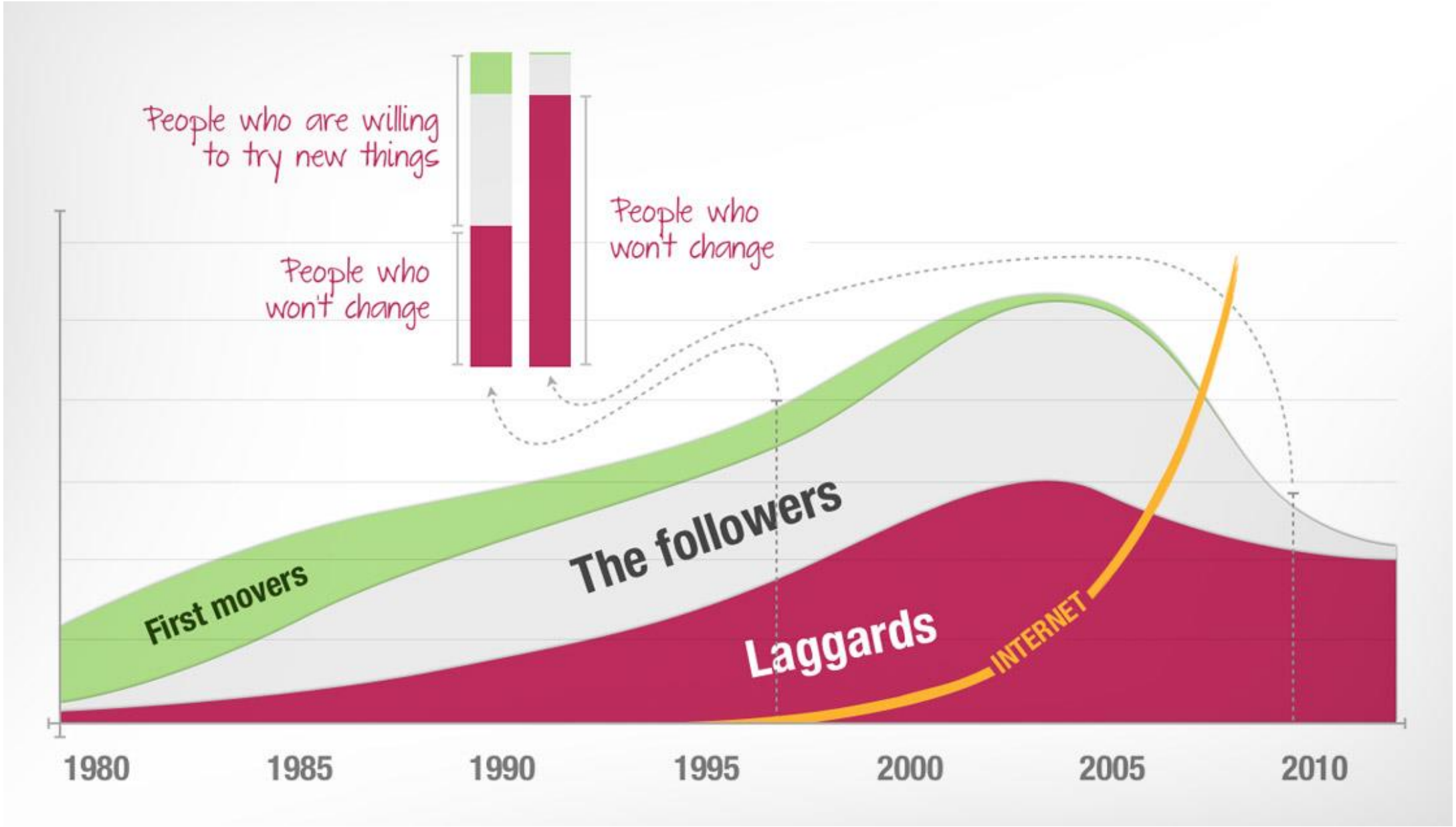
Swedish launch funded by VINNOVA Sweden's Innovation Agency

How to learn more about AI?

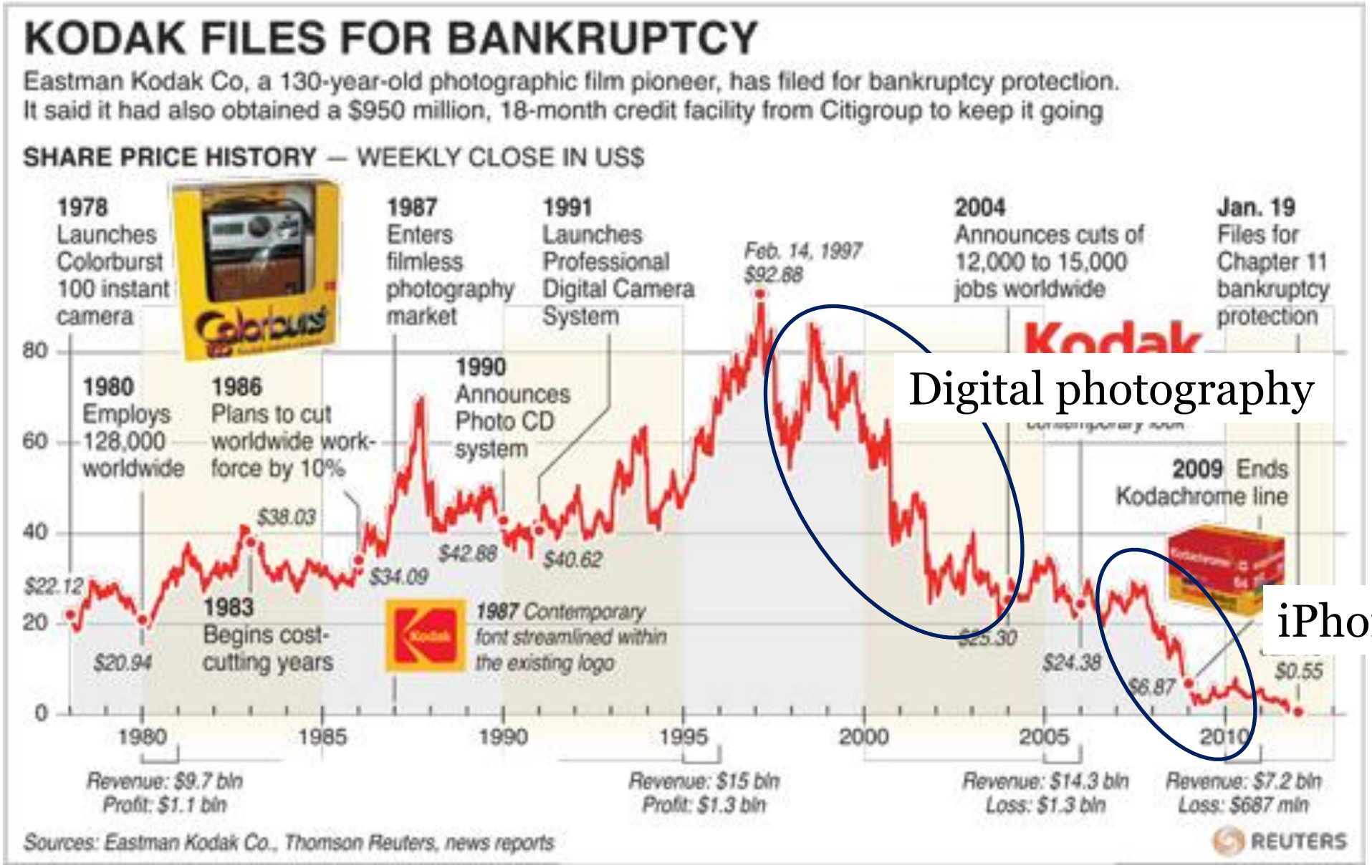
- Watch some overview presentations and read some popular science pieces
- Take an overview online course, e.g. https://www.elementsofai.com/
- Take a more technical online course and do some programming/problem solving, e.g. https://developers.google.com/machine-learning/crash-course/
- Take a real problem from your organization and try to solve it using AI-techniques
- Take a technical course, either online or at a university
- Build real solutions



https://www.baekdal.com/trends/understanding-change-and-predicting-trends/



https://www.baekdal.com/trends/understanding-change-and-predicting-trends/



http://www.fpa-trends.com/article/how-manage-forecasting-risk

Take Away Message

- AI is about understanding intelligence and develop systems that exhibit intelligent behavior. AI will affect all aspects of our society.
- **Trust is essential**.
- To be **trustworthy** an **AI-system** should be **legal**, **ethical** and **robust**.
- Education and life long learning will be absolutely necessary.
  - **Take the online course Elements of AI (http://elementsofai.se)**
- Recommendations
  - **Learn** more! Experiment!
  - **Do** concrete projects on important topics.
  - **Scale up**.
- ***Digital tools will only provide value when you learn how to use them effectively and adapt your organization to leverage them!***
- **Human + AI**


Respect for human autonomy


Prevention of harm


Fairness


Explicability

Trustworthy Human-Centric AI

Fredrik Heintz, Dept. of Computer Science
Linköpings universitet
fredrik.heintz@liu.se
@FredrikHeintz

LINKÖPINGS UNIVERSITET